The Dissertation Committee for Yaelim Lee Certifies that this is the approved version of the following dissertation:

TRANSCRIPTION FACTOR NETWORKS AND CHROMATIN REMODELER FUNCTION IN GENE REGULATION ON THE EUKARYOTIC GENOME

Committee:

Vishwanath R. Iyer, Supervisor

Edward M. Marcotte

Claus O. Wilke

Arlen W. Johnson

Jonghwan Kim

TRANSCRIPTION FACTOR NETWORKS AND CHROMATIN REMODELER FUNCTION IN GENE REGULATION ON THE EUKARYOTIC GENOME

by

Yaelim Lee, B.S.; M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2016

Dedication

To my family.

Acknowledgements

I would like to thank my advisor, Dr. Vishy Iyer for his endless patience throughout my graduate study. I feel extremely lucky to have him as my advisor. He has always respected my pace, no matter how slow. I would not have enjoyed learning and progressed this far without his trust and guidance.

I would also like to thank my committee members, Dr. Edward Marcotte, Dr. Claus Wilke, Dr. Arlen Johnson, and Dr. Jonghwan Kim for taking the time to help me throughout my PhD. I am also grateful to Dr. Claus Wilke and Dr. Bindu Viswanathan for providing me with an opportunity to work with as a rotation student and a teaching assistant. It was this experience that piqued my interest in bioinformatics and biostatistics.

I would like to thank my colleagues, Bum-Kyu, Daechan, Haridha, Anna, Amelia, Yunyun and Dia for their positive interactions that not only helped me progress through research but also made each day in the lab joyful. Their intellectual stimulation as well as kind encouragement enabled me to remain persistent throughout 6 years.

Finally, I would like to thank my parents who came across the Pacific multiple times to support my work, and to my husband and son whose dependable presence and joyous laughter helped me get through all the ups and downs in my PhD.

TRANSCRIPTION FACTOR NETWORKS AND CHROMATIN REMODELER FUNCTION IN GENE REGULATION ON THE EUKARYOTIC GENOME

Yaelim Lee, Ph.D.

The University of Texas at Austin, 2016

Supervisor: Vishwanath R. Iyer

Numerous events – from histone modification and transcription factor binding to gene expression – take place on eukaryotic chromatin, while cells are constantly exposed to dynamic stimuli ranging from spatial and temporal cues to environmental and extracellular signals. The cell's ability to respond and adjust accordingly is directly related to cell fitness and viability. With the advent of next-generation sequencing, investigating these events has been enabled at nucleotide resolution but across the entire genome. In this dissertation, I investigate changes on eukaryotic genomes including yeast and human, which are triggered by stress and by loss of a protein of interest, by analyzing genomics data generated mainly through next-generation sequencing. In Chapter 1, I determine how yeast cells achieve transcriptional reprogramming in response to heat stress by first identifying the complete set of transcription factors that are essential for heat stress conditions. This is further explored by identifying both the target loci bound by the transcription factors under conditions of heat-stress, as well as the genes that

require the function of the transcription factor for normal transcriptional response to heat stress. In Chapter 2, I study a chromatin remodeling factor, *CHD1* (Chromodomain Helicase DNA binding protein 1) with regard to two aspects: first, what factors provide specificity for Chd1 positioning on chromatin, by examining the role of proteins that physically or genetically interact with Chd1, and second, what is the relationship of *CHD1* with the hallmark of chromatin modifications, histone H3 tri-methylation at Lys 4 and Lys 36, by investigating changes in these histone methylation marks in the absence of *CHD1*. Additionally, I show a novel functional link between *CHD1* and RNA splicing through analysis of intron retention in transcripts produced in the *CHD1* mutant. Lastly, I investigate *CHD1* role in human glioblastoma cell line by generating a *CHD1* knock out via the CRISPR/Cas9 genome editing system. Taken together, the work presented in this dissertation provides novel approaches, discoveries, and intriguing insights into how eukaryotic chromatin experiences dynamic alterations in response to various perturbations on a genome-wide scale.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

1.1 Background

1.1.1 TRANSCRIPTION ON EUKARYOTIC CHROMATIN

Chromatin is a complex of DNA and proteins that forms chromosomes within the nucleus of eukaryotic cells. The basic structural unit of chromatin is the nucleosome which is composed of two each of four canonical histones (H2A, H2B, H3, and H4) and approximately 150 bps of DNA wrapping 1.7 turns around the histone octamer (Annunziato, 2008).

For transcription initiation, multiple proteins are recruited on chromatin to form a preinitiation complex, which include *cis*-regulatory element (CRE)-binding TFs, coactivators such as chromatin remodelers and histone modifying enzymes, the mediator complex, general TFs (GTFs), and RNA polymerase II (RNAPII) (Figure 1.1). Once the preinitiation complex has formed, RNAPII starts to move from the transcription start site (TSS) along the DNA strands, synthesizing RNA. This process is called transcription (Fromm et al., 2013) (Figure 1.1).

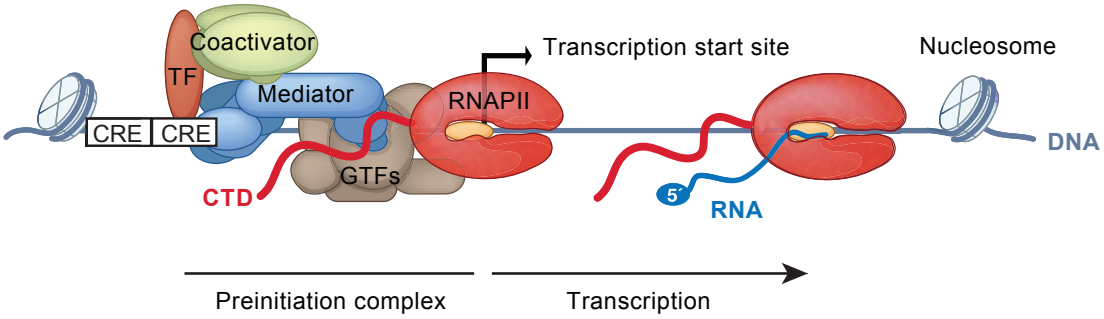


Figure 1.1: Chromatin, transcription factors, and transcription[i]. *Notes* – CRE: cis-regulatory elements, TF: transcription factor, GTFs: general transcription factors, RNAPII: RNA polymerase II, CTD: RNAPII C-terminal domain.

The transcription process can be further subdivided into early elongation and productive elongation stages based on the rate of synthesizing RNA by RNAPII. In metazoans, RNAPII often pauses its passage shortly after transcription initiation (*e.g.*, after synthesis of 20 ~ 60 nucleotide-long RNA) (Liu et al., 2015), and this is considered a quality checkpoint for the nascent transcript as well as a way to ensure that RNAPII is competent for productive elongation (Jonkers and Lis, 2015). This stage, from the transcription initiation to the pause, defines the early elongation phase (Jonkers and Lis, 2015; Liu et al., 2015), and during it RNAPII is distinguished by phosphorylation at serine 5 of its C-terminal domain (CTD). After RNAPII is released from the pause, it enters the productive elongation stage, and RNAPII is then phosphorylated at serine 2 of its CTD in this phase (Jonkers and Lis, 2015). While yeast do not exhibit the same type of RNAPII pausing signatures, except maybe at 3' splice sites for intron containing genes

[i] Modified from Fromm, G., Gilchrist, D.A., and Adelman, K. (2013). SnapShot: Transcription regulation: pausing. Cell *153*, 930-930 e931.

(Alexander et al., 2010), they do exhibit the same shift in phosphorylation on their RNAPII CTD domains. Their transcription can, therefore, also be characterized as experiencing to distinct phases.

These distinct phosphorylation marks in the RNAPII CTD domain serve as docking sites for interacting with other proteins in a stage-specific manner, including the mediator complex, co-transcriptional pre-mRNA processing factors such as RNA capping enzymes and RNA splicing factors, elongation factors, termination factors, as well as histone modifying enzymes (Hajheidari et al., 2013; Kim et al., 2010; Phatnani and Greenleaf, 2006).

1.1.2 CHROMATIN REMODELERS

The repeating nucleosomes comprising chromatin are considered to be a barrier that RNAPII has to overcome for transcription. Thus, chromatin experiences dynamic structural alterations during transcription, not only by loosening its compact conformation to allow RNAPII to progress for RNA synthesis, but also by returning to its repressive structure with well-positioned nucleosomes in order to prevent spurious transcription initiation in the wake of RNAPII. Chromatin remodelers come into play in producing these alterations on chromatin by replacing canonical histones with histone variants, and by sliding, spacing, evicting, and assembling nucleosomes to facilitate regulation of the transcription process (Petty and Pillus, 2013).

There are four families of chromatin remodeler complexes, which are defined by similarities between their protein domains: SWI/SNF, ISWI, INO80, and CHD families (Table 1.1) (Clapier and Cairns, 2009).

Table 1.1: Chromatin remodeler families and functions. *Chromatin remodeling complexes are found in yeast.

Families	Functions	Complexes*
SWI/SNF	Sliding and evicting nucleosomes	SWI/SNF, RSC
ISWI	Assembling and spacing nucleosomes	ISW1a, ISW1b, ISW2
INO80	Replacing H2A with the variant H2A.Z	INO80, SWR1
CHD	Sliding, evicting, spacing, and assembling nucleosomes	CHD1

1.1.3 Histone Modifications in Transcription

Reversible covalent modification of histones in nucleosomes, such as (de)acetylation and (de)methylation, has important roles in regulating transcription by altering chromatin structure and by providing binding specificity to interacting proteins (Kurdistani and Grunstein, 2003; Martin and Zhang, 2005).

Acetylation and methylation on lysine residues in the core histone H3 have been extensively studied and display a close relationship with transcription. Table 1.2 shows a summary of histone H3 modifying enzymes that have been found in yeast.

Table 1.2: Histone H3 modifying enzymes found in yeast[i].

Histone modifying enzymatic activity	Histone H3 modifying enzymes (Lysine residues)
Histone acetyltransferase activity	Gcn5 (K9, K14, K18, K23, K27), Rtt109 (K9, K27, K56)
Histone deacetylase activity	Hda1, Hos1, Hos2, Hos3, Rpd3
Histone methyltransferase activity	Set1 (K4), Set2 (K36), Dot1 (K79)
Histone demethylase activity	Jhd1 (K36), Jhd2 (K4), Gis1 (K36), Rph1 (K9, K36)

A general outcome of histone acetylation is the formation of a more relaxed chromatin structure via attenuation of the electrostatic interactions between negatively charged DNA and positively charged histones, which leads to permissive conditions for transcription. On the other hand, histone deacetylation – the removal of acetyl groups from histone lysine residues – is involved in transcriptional repression (Kurdistani and Grunstein, 2003).

Histone lysine methylation is considered a key determinant in defining distinct chromatin states, such as promoter states, transcribed states, regions of active transcription, regions of repressed transcription, and repetitive elements (Black et al., 2012). For example, tri-methylation of lysine 4 of histone H3 (H3K4me3) is a hallmark of the transcription start sites of genes, while tri-methylation of lysine 36 of histone H3

[i] Information has been collected based on annotations in Saccharomyces Genome Database (www.yeastgenome.org).

(H3K36me3) is found over gene bodies and is strongly correlated with transcription level (Black et al., 2012). Based on their positions at genes as well as based on their interacting proteins, which are considered effectors of these modifications, roles of H3K4me3 and H3K36me3 have been uncovered in the regulation of transcription initiation and elongation, respectively (Black et al., 2012; Krogan et al., 2003; Shilatifard, 2012).

1.2 Research Overview

To understand gene function, genes need to be studied under conditions where their functions are known to be required. This is because gene expression is dynamic, and changes upon recognition of specific cues, ranging from spatial and temporal signals to environmental and extracellular ones. In addition, a gene needs to be examined as a member of a complex network because it does not function alone. Instead, it interacts with partners whose functions have regulatory roles or are required for coordinated effects, such as TFs and chromatin remodelers. For these reasons, a large amount of research has been directed towards constructing a transcription factor network (also called a transcription regulatory network) and to elucidating causality between TFs and regulatory targets on a genome-wide scale in yeast. However, several limitations have prevented this field of study from moving forward.

The first limitation is that experimentally constructed transcription factor network in previous studies such as Lee *et al*. (2002), Harbison *et al*. (2004), and Hu *et al*. (2007) rely exclusively on one of two techniques: either chromatin immunoprecipitation with

microarray (ChIP-chip) or microarray gene expression profiling (Harbison et al., 2004; Hu et al., 2007; Lee et al., 2002). The ChIP-chip assay provides the binding loci of TFs across the genome, which enables us to infer direct binding target genes of the TFs (Harbison et al., 2004; Lee et al., 2002). Microarray gene expression profiling in TF-deletion strains demonstrates differentially expressed genes that are under control of a particular TF, and which are functionally regulated target genes (Hu et al., 2007). Although these two types of datasets play a complementary role in specifying causality between TFs and target genes within a network, the datasets have been generated from independent studies making integration difficult due to limited overlap (Hu et al., 2007).

An additional complication is that because these transcription factor networks were constructed under normal growth conditions, they lack the data under stress conditions, where the function of a particular TF is required. Indeed, two seminal studies from the Ideker Lab investigating how a eukaryotic cell reprograms its regulatory interactions (between TF and gene) and genetic interactions (between gene-gene) in response to DNA damage show that both types of interactions are completely changed under conditions that induce DNA damage (Bandyopadhyay et al., 2010; Workman et al., 2006). This emphasizes the significance of probing the dynamic interactions that enable cells to survive and thrive in varying environmental and genetic contexts. In this regard, the transcription factor network, which I attempt to construct in Chapter 2 by using both ChIP sequencing and RNA sequencing/microarray techniques under static and perturbed conditions, provides insight into condition-specific transcriptional regulatory mechanisms underlying the heat shock response, which is one of the most conserved stress responses

(Richter et al., 2010). I expect that the data from this project will serve as a landmark dataset valuable in its own right, similar to previous studies such as Lee *et al*. (2002), Harbison *et al*. (2004), and Hu *et al*. (2007), which have been extensively used and reanalyzed in dozens of subsequent publications from unaffiliated labs.

Moreover, a genome-wide transcription factor network has potential for revealing novel genes, TFs, and interactions, which are functionally associated. From this perspective, the network that I construct will not only provide many testable regulatory interactions, but also allow me to make predictions of gene function that were previously unidentified. Transcriptional regulatory mechanisms are highly conserved from yeast through humans; therefore, genuine interactions and predictions derived from the network will provide a fundamental framework for identifying and testing important components of transcriptional regulation in higher organisms including human.

In Chapter 3, I study one of the universally conserved chromatin remodelers, the CHD family (Table 1.1). Yeast has one CHD-type remodeler, *CHD1* (Table 1.1), and *CHD1* mutants exhibit a high degree of aberrant nucleosome structure. In Chapter 3, I attempt to answer three questions: i) How is Chd1 recruited to chromatin; ii) What is the relationship of Chd1 with two histone modifications, H3K4me3 and H3K36me3; and iii) What is the functional impact of Chd1?

Our study provides information on i) the actions of ten key proteins interacting with Chd1, and their role in Chd1 recruitment; ii) a clear link between Chd1 and two histone methylation marks; iii) evidence to support a role for Chd1 in RNA splicing.

CHAPTER 2

TRANSCRIPTIONAL CONTROLS UNDER STRESS CONDITIONS

2.1 Introduction

Cells are constantly exposed to dynamic stimuli ranging from spatial and temporal cues to environmental and extracellular signals. The cell's ability to control gene expression levels by transcriptional regulation is a primary mechanism ensuring the cell is able to perform its normal functions in a condition-specific manner. For this reason, precise and accurate regulation of transcription is directly related to cell fitness and viability.

Transcriptional regulation involves a complex interplay between regulatory proteins called transcription factors (TFs) and regulatory DNA sequences called promoters. In eukaryotes, DNA is tightly wrapped into a higher order structure called chromatin, which mediates the interaction of TFs with promoters. Chromatin remodeling factors, including histone modifiers, regulate chromatin structure by loosening or tightening chromatin domains via covalent modifications on histones and replacement with histone variants, or by providing a binding site for additional TF recruitment.

Strikingly, transcriptional regulatory mechanisms are strongly conserved from yeast to humans. Indeed, all the regulatory proteins found in yeast, including TFs and

chromatin modifiers, have counterparts in human cells. One of the most highly conserved transcriptional regulatory programs is the heat shock response (Akerfelt et al., 2010; Westerheide and Morimoto, 2005), where transcriptional regulation is mediated by the binding of heat shock transcription factors (HSFs) to DNA promoter elements called heat shock elements, which can be found across the genome (Hahn et al., 2004). Through prompt and coordinated transcriptional responses, cells adapt efficiently to altered temperatures and maintain cellular homeostasis without accumulating defects that may lead to critical cell damage (Richter et al., 2010). It is believed that the transcriptional responses mediated through the activation of HSFs are triggered by unfolded or misfolded proteins, which can fail to fold properly due to high temperatures (Richter et al., 2010). Aberrant proteins can be produced not only by heat shock but also by a variety of other stressors such as oxidative stress and toxic substances (Richter et al., 2010); therefore, understanding genome-wide transcriptional changes that are triggered in response to heat shock would provide insight into general and fundamental principles of transcriptional control under stress conditions.

A large amount of genomic research has investigated what fraction of the genome responds to stress. This includes quantitatively measuring gene expression levels following heat shock, which reveals that the expression of approximately 20% of all genes in the yeast genome is significantly altered by heat shock (Causton et al., 2001; Gasch et al., 2000; Zanton and Pugh, 2004). Although several subsequent studies have focused on trying to explain these observed transcriptional changes, key stress-response TFs in yeast (Hsf1, Msn2 and Msn4) account for only a subset of the observed changes in

gene expression (Gasch et al., 2000) (*e.g.*, only 3% of all yeast genes are targeted by Hsf1 following heat shock (Hahn et al., 2004)). This strongly suggests that there are many more transcriptional regulators that promote transcriptional changes in response to heat shock.

A transcription factor network is a representation of the relationship between transcriptional regulators and target genes. Because a transcription factor network can provide a systematic view of transcriptional regulation by explaining the causality between the TFs and target genes and the hierarchy among the TFs (Jothi et al., 2009), several studies have put a large amount of effort into experimentally creating transcription factor networks in yeast by identifying genome-wide binding sites and target genes of tens to hundreds of TFs (Harbison et al., 2004; Hu et al., 2007; Lenstra et al., 2011; Venters et al., 2011; Workman et al., 2006). These studies, except for that by Workman *et al*. (Workman et al., 2006), relied exclusively on one of two techniques to determine the regulatory targets: either chromatin immunoprecipitation with microarray (ChIP-on-chip) or microarray gene expression profiling. The ChIP-on-chip assay provides binding locations of TFs across the genome, which enables identification of target genes directly bound by the TFs (Harbison et al., 2004; Venters et al., 2011). However, the ChIP-based network has a limitation that TF binding itself does not guarantee that the TF controls expression of the target gene assigned based on the binding event.

Microarray gene expression profiling in TF deletion strains demonstrates differentially expressed genes that are under control of a particular TF (Hu et al., 2007).

With regard to the expressional change, these target genes are functionally associated with the TF. Yet, the microarray-based network is limited for solving the hierarchy of target genes because there is no further information provided for understanding the interrelationships between the differentially expressed genes in the absence of the TF.

Workman *et al*. (Workman et al., 2006) attempted to understand how yeast cells reprogram the regulatory relationships between TFs and target genes in response to DNA damage by using both ChIP-on-chip and microarray gene expression profiling. The results show that the relationships between TFs and target genes are completely changed under conditions that induce DNA damage. This result emphasizes the significance of probing the dynamic transcriptional regulation that enables cells to survive and thrive in varying environmental contexts.

In Chapter 2, first, we exhaustively identified TFs and chromatin remodeling factors necessary for normal growth and survival of cells under heat stress conditions in yeast. Next, we systematically characterized a set of target loci bound by the TFs with and without heat shock by using chromatin immunoprecipitation (ChIP) sequencing. Additionally, we generated gene expression profiles with the TF deletion mutants upon heat shock via microarray or RNA sequencing to define functional links between target transcripts and the TF.

From the large-scale data that we generated, we discovered not only heat shock-induced binding loci of the TFs but also genes whose transcriptional response upon heat shock are under control of the TFs. Interestingly, we identified a novel Hsf1 binding site at an intergenic region where no heat shock-induced gene expression was observed. Our

RNA sequencing data revealed that activation of a non-coding transcript adjacent to the Hsf1 binding. This result indicates functionality of TF binding for regulating transcription beyond coding transcripts (*i.e*, genes) in response to heat stress.

All in all, the data presented in Chapter 2 aimed to provide informative resources for understanding not only a global view of transcriptional regulation in response to heat shock but also detailed mechanisms with regard to the contributions of individual TFs.

2.2 Materials and Methods

2.2.1 Yeast as a Model Eukaryotic System

We used the yeast *Sacchromyces cerevisiae*, which is a suitable model organism for systems-based approaches, because it has the following two attributes: i) availability of a well-annotated, sequenced genome, and ii) availability of technical resources such as single gene knockout strains (*i.e.*, Yeast Knockout Collection) and epitope-tagged strains (*i.e.*, Yeast TAP Tagged ORF Collection), enabling the easy manipulation of individual TFs and chromatin remodelers in order to elucidate their roles in heat stress response. We screened approximately 400 individual deletion strains in order to identify essential TFs under heat stress conditions. Table 2.1 shows the yeast strains that we used for data collection for Section 2.3.2 and Section 2.3.4, following a post-screening prioritization process (described in Section 2.3.1). For wild-type (WT) yeast strain, we used the S288C-derivative laboratory strain, BY4741 (*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*).

Table 2.1: Yeast strains used in Chapter 2. *Notes* – NA: Not available in the Yeast Knockout Collection. *Aneuploidy and partial deletion were caveats that we found in the Yeast Knockout Collection strains. In theses cases, we created the strains via yeast transformation (described in Section 3.2.1).

Transcription Factors	**Yeast Knockout Collection**	**Yeast TAP Tagged ORF Collection**	**Strains we created**
Hsf1	NA	NA	*HSF1-TAP*
Msn2		*MSN2-TAP*	*msn2Δmsn4Δ*
Msn4		NA	*MSN4-TAP*
Gcr1	NA	*GCR1-TAP*	
Gcr2	*gcr2Δ*		
Swi6	Aneuploidy*		*swi6Δ*
Cyc8	*cyc8Δ*	*CYC8-TAP*	
Tup1	*tup1Δ*	*TUP1-TAP*	
Ume6	Aneuploidy*	*UME6-TAP*	
Sua7	NA	*SUA7-TAP*	
Paf1	Aneuploidy*	*PAF1-TAP*	*paf1Δ*
Leo1		*LEO1-TAP*	
Rpb9	*rpb9Δ*	*RPB9-TAP*	
Spt4	*spt4Δ*	*SPT4-TAP*	
Spt5	NA	*SPT5-TAP*	
Mbf1	Partial deletion*	*MBF1-TAP*	*mbf1Δ*
Sin4	*sin4Δ*	*SIN4-TAP*	
Srb2	*srb2Δ*	*SRB2-TAP*	
Spt2	*spt2Δ*	*SPT2-TAP*	
Spt16	NA	*SPT16-TAP*	
Ino80		*INO80-TAP*	
Arp8	*arp8Δ*	*ARP8-TAP*	
Swr1		*SWR1-TAP*	
Htz1		*HTZ1-TAP*	

2.2.2. GENOMICS METHODOLOGIES

2.2.2.1 Chromatin Immunoprecipitation (ChIP) Sequencing

A transcription regulator binds to specific genomic loci such as promoters, intergenic regions (*i.e.*, gene body), or intragenic regions to control transcription. In order to determine the physical binding loci of TFs across the yeast genome, we used ChIP sequencing, which allows for the determination of the DNA fragments bound by a specific TF with single-nucleotide resolution through high-throughput sequencing (Park, 2009). For this purpose, we took advantage of the Yeast TAP tagged ORF Collection, where each gene is tagged with a high-affinity epitope (TAP-tag) and expressed from its natural genomic location (Ghaemmaghami et al., 2003).

For the ChIP experiments, first, proteins were cross-linked to DNA by adding formaldehyde and the cross-linking reaction was then quenched with glycine. The yeast cells were lysed with lysis buffer and then agitated with glass beads using a bead beater. Next, we used a sonicator to shear the protein-bound genomic DNA in order to facilitate the subsequent immunoprecipitation reaction. To immunoprecipitate the TF of interest, we used IgG conjugated sepharose beads that bind to the TAP-tag with high specificity. Subsequently, the isolated DNA and protein complexes were reverse crosslinked, the DNA was purified, and barcoded DNA sequencing libraries were prepared. Lastly, sequencing was done through the Illumina HiSeq System.

2.2.2.2 Microarray and RNA Sequencing

TF-binding, by itself, does not reveal whether a transcriptional regulator participates in the activation or repression of gene expression. To identify transcripts which are differentially expressed upon heat stimulus and to determine whether changes in expression stem from the role of a specific regulator under this condition, we carried out either microarray based gene expression profiling or RNA sequencing using TF deletion mutants with and without heat stimulus.

For the microarray experiments, the first step was to isolate total RNA from the yeast cells. Subsequently, polyadenylated RNA was isolated by reverse transcription with an oligo-dT primer, and then the complementary strand was synthesized to make double-stranded DNA (dsDNA). The dsDNA was fluorescently labeled with the Cy3 dye and hybridized to the NimbleGen *S. cerevisiae* Gene Expression Array. Following washing and scanning the microarrays, we extracted expression values for individual genes in the yeast genome.

For the RNA sequencing experiments, we isolated total RNA from the yeast cells and incubated the total RNA with magnetic beads conjugated to oligo-dT to enrich poly-adenylated RNA. Next, a sequencing library with the poly-A RNA was prepared, and it was sequenced through the Illumina HiSeq System.

2.2.2.3 Data Acquisition

Table 2.2 shows the data sets that we collected for this study.

Table 2.2: Summary of experimental data sets investigated in Chapter 2. *Notes* – NA: not applicable. N and HS: normal and heat shock conditions, respectively. RNAPII CTD S2p: RNA polymerase II C-terminal domain serine 2 phosphorylation. RNAPII CTD S5p: RNA polymerase II C-terminal domain serine 5 phosphorylation. H3: histone H3. More details about the immunoprecipitants are described in Appendix A.

Strain	**Experiment**	**Immunoprecipitant**	**Conditions, Replicates #**
HSF1-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 2
MSN2-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
MSN4-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
GCR1-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
SWI6-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
CYC8-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
TUP1-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
UME6-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
SUA7-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
PAF1-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
LEO1-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
RPB9-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
SPT4-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
SPT5-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
MBF1-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
SIN4-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
SRB2-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
SPT2-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
SPT16-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
INO80-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
ARP8-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
SWR1-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 1
HTZ1-TAP	ChIP-Seq	IgG sepharose beads	N and HS, 2
WT	ChIP-Seq	Anti-RNAPII CTD S5p	N and HS, 1

(Table 2.2 continued.)

Strain	Experiment	Immunoprecipitant	Conditions, Replicates #
WT	ChIP-Seq	Anti-RNAPII CTD S2p	N and HS, 1
WT	ChIP-Seq	Anti-H3	N and HS, 1
WT	ChIP-Seq	IgG sepharose beads	N and HS, 3
msn2Δmsn4Δ	RNA-Seq	NA	N and HS, 2
swi6Δ	RNA-Seq	NA	N and HS, 2
tup1Δ	RNA-Seq	NA	N and HS, 2
WT	RNA-Seq	NA	N and HS, 2
rpb9Δ	Microarray	NA	N and HS, 2
gcr2Δ	Microarray	NA	N and HS, 2
arp8Δ	Microarray	NA	N and HS, 2
spt2Δ	Microarray	NA	N and HS, 2
srb2Δ	Microarray	NA	N and HS, 2
sin4Δ	Microarray	NA	N and HS, 2
mbf1Δ	Microarray	NA	N and HS, 2
paf1Δ	Microarray	NA	N and HS, 2
spt4Δ	Microarray	NA	N and HS, 2
cyc8Δ	Microarray	NA	N and HS, 2
WT	Microarray	NA	N and HS, 4

2.3 Results

2.3.1 Essential Transcription Factors Under Heat Stress Condition

If a TF is essential under the heat stress conditions, its loss-of-function mutation would show a visible phenotype at non-optimal high temperature. We compiled a broad set of ~400 transcriptional factors (TFs) whose annotation is involved in the transcriptional regulation through DNA sequence-specific binding or chromatin structure remodeling in Saccharomyces Genome Database (SGD) and Munich Information Center for Protein Sequences (MIPS). Additionally, we added putative DNA binding proteins that possibly affect transcription, which are uncovered by Hall *et al*. (2004) (Hall et al., 2004), to the TF candidates. To identify the TFs that are required for growth or survival under the heat stress conditions, we performed a phenotypic growth screen for the TF candidate mutants, and monitored if the mutants show growth defects under heat stress conditions (Figure 2.1).

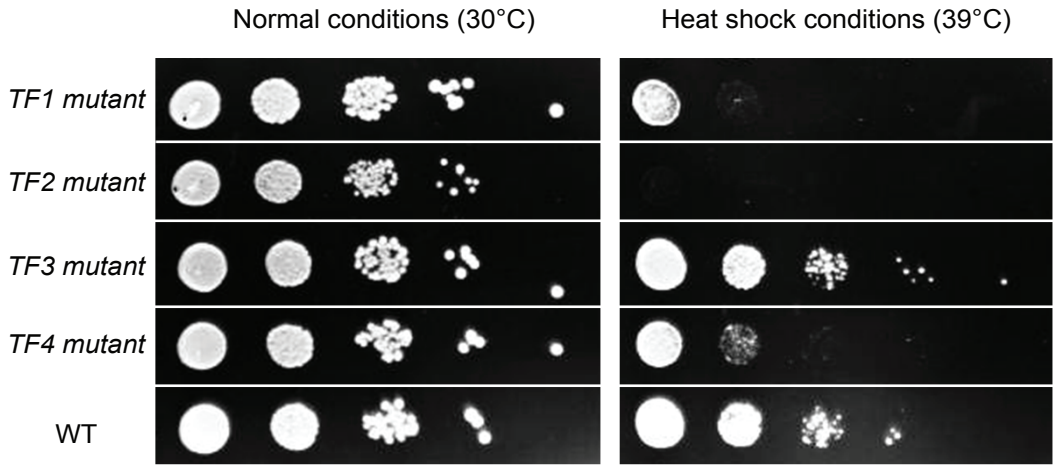


Figure 2.1: A schematic representation of the growth screening. Each of the TF-deletion mutants is spotted onto two YPD plates with five serial dilutions ranging from the O.D600 of 0.1 to 0.00001. One plate is incubated under normal conditions and the other plate is incubated under the heat stress conditions, at 30ºC and 39ºC, respectively. The growth of the TF mutants are monitored to examine whether individual TF mutants show visible phenotypes such as lethality (*e.g.*, *TF2 mutant* here), noticeable growth defects (*e.g.*, *TF1* and *TF4 mutants* here), and slow growth under heat stress conditions.

As a result, 49 TF deletion mutants showed phenotypic defects with a varied degree under heat stress conditions in at least two biological replicates. The TFs are summarized in Figure 2.2 according to the severity of the defects, from the no survival to the slow growth categories.

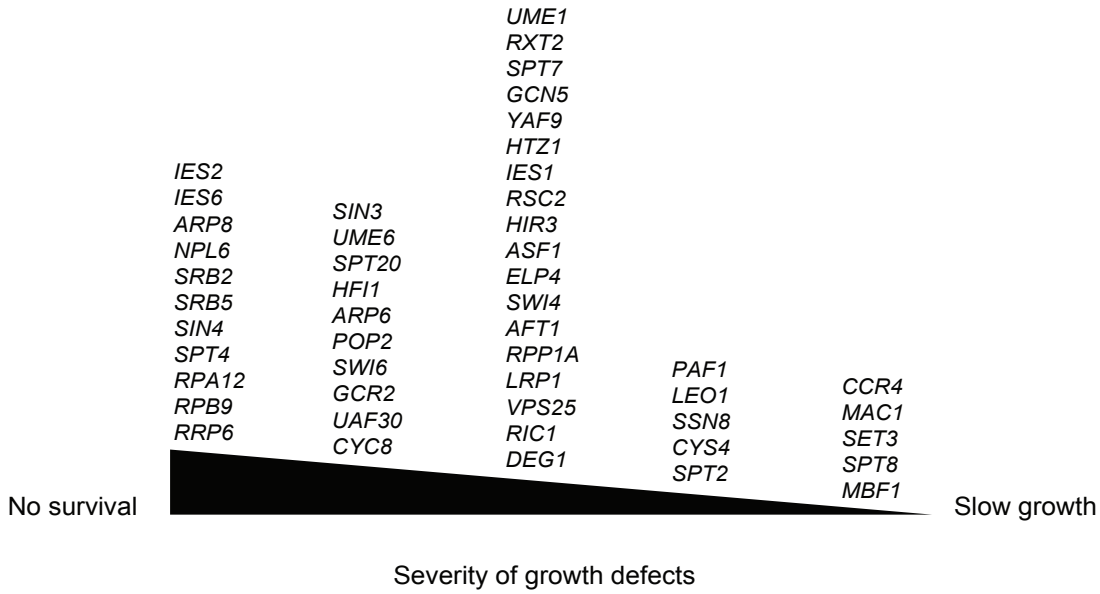


Figure 2.2: 49 TFs showing growth defects upon loss of the TF under heat stress conditions. The TFs are listed according to severity of the growth defect on the growth screening. The left most TF list shows no growth, and the middle TF list shows the first or/and the second serial dilution(s) on the YPD plate at 39ºC. The right most TF list shows a comparable growth to WT eventually (at 72 hours of incubation) but the examination at earlier time points (*e.g.*, 24 hours and 48 hours of incubation) shows the delayed appearance of serial dilutions, when compared to WT.

One potential issue with the yeast deletion library is a possibility of errors arising during the generation of the deletion strains, such as accumulation of cryptic mutations that might affect yeast fitness, or handling and copying of the library resulting in misidentification or cross-contamination among strains. Indeed, Lenstra *et al*. (2011) (Lenstra et al., 2011) found 26 out of 165 deletion library strains used for their study have different issues, such as moderate expression of the deleted gene, aneuploidy, and spurious mutations. Also, we found two cases of aneuploidy and one case of the gene partially deleted (Table 2.1) from the initial gene expression profiles. To rule out these

issues inherent in the deletion collection, we created TF deletion mutants for the 49 TFs by transformation (described in Section 3.2.1), and defined them as remade strains. Subsequently, we carried out the growth assay with the remade TF deletion strains to recapitulate the growth defects that we observed from the deletion library strains. We found an inconsistency in the phenotypes under heat stress conditions between the deletion library and the remade strains for four TFs – *UME1*, *CCR4*, *AFT1* and *CYS4*. We, therefore, dropped these four TFs from our candidate list. In summary, we identified 45 TFs that are necessary for the normal growth under heat stress conditions in yeast.

Among the 45 TFs identified, we noticed that several TFs constitute the same protein complex. For example, Ies1, Ies2, Ies6, and Arp8 are members of a chromatin remodeling INO80 complex (Figure 2.3). Accordingly, we categorized the 45 TFs based on the protein complex and biological process, resulting in 5 broad processes and 12 complexes (Table 2.3). For the following target identification of the TF (Section 2.3.2 and Section 2.3.4), we prioritized the TFs according to whether the TF has DNA binding activity or DNA binding domain, which is critical for the success on ChIP experiments, attempting to maximize the number of protein complexes. The prioritized TFs are shown in red in Table 2.3. In addition to this list, we included important TFs for transcription regulation in response to heat shock, such as Hsf1, Msn2, and Msn4. Also, essential transcription initiation and elongation factors, such as Sua7, Spt5, and Spt16, were added to the final list.

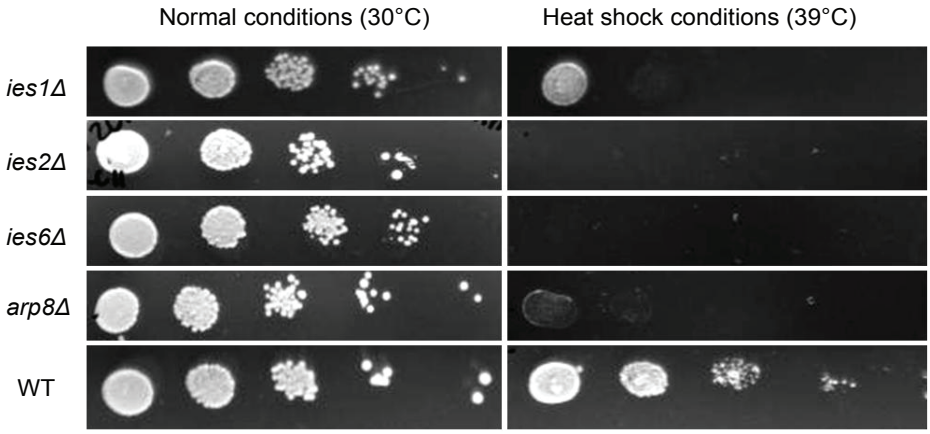


Figure 2.3: Growth defects of the INO80 complex components, including Ies1, Ies2, Ies6 and Arp8, under heat stress conditions.

Table 2.3: Categorizing the TFs that are sensitive to heat stress depending on the protein complex and biological process. The prioritized TFs for identifying regulatory targets (Section 2.3.2 and Section 2.3.4) are colored in red.

Biological process	**Protein complex**	**Transcription factors**
Histone modification	RPD3L complex	*SIN3*, *UME6*, *RXT2*
	SAGA complex	*SPT2*, *SPT7*, *SPT8*, *SPT20*, *HFI1*, *GCN5*
	SET3 complex	*SET3*
Chromatin remodeling	SWR1 complex	*ARP6*, *YAF9*
	INO80 complex	*IES1*, *IES2*, *IES6*, *ARP8*
	RSC complex	*RSC2*, *NPL6*
Nucleosome assembly	HIR complex	*HIR3*
		ASF1, *HTZ1*
RNAPII-associated	RNAPII	*RPB9*
	PAF1 complex	*PAF1*, *LEO1*
	Mediator complex	*SRB2*, *SRB5*, *SIN4*, *SSN8*
	DSIF complex	*SPT4*
	Elongator complex	*ELP4*
Transcription cofactor		*SWI4*, *SWI6*, *GCR2*, *MAC1*, *CYC8*, *MBF1*
RNAPI-associated		*RPA12*, *UAF30*
Others		*POP2*, *LRP1*, *RRP6*, *RPP1A*, *VPS25*, *RIC1*, *DEG1*

2.3.2 Genome-wide Localization of Transcription Factors upon Heat Shock

If occupancy of a TF at specific genomic locations is significantly higher or lower upon heat stress, then the alteration is likely to be induced by heat shock to regulate transcription of the adjacent region.

We generated 66 ChIP sequencing data sets, including 23 TFs, 2 RNAPII CTD phosphorylation marks, and 3 mock controls, in which we carried out immunoprecipitation with IgG beads in cells expressing no TAP-tagged protein, both in the presence and absence of heat shock (summarized in Table 2.2). After mapping of the ChIP sequencing reads to the yeast reference genome (SacCer3) using the aligner BWA (Li and Durbin, 2009), we visualized the data on a local mirror of the UCSC genome browser to see TF binding events (Figure 2.4 and Figure 2.5). We found that TF binding events on the yeast genome are highly dynamic in response to heat stress. For example, *HSP104* and *SSA2* are the heat shock protein genes whose expression is activated in response to heat shock. As anticipated, we found high occupancy of several TFs induced upon heat stimulus at the locus including two genes (Figure 2.5). Also, we were able to distinguish between TFs with local binding (narrow peak TFs, *e.g.*, Hsf1 and Sua7) and TFs showing widespread binding across the entire transcript (broad peak TFs, *e.g.*, Spt2 and Spt5). These broad peak TFs were a part of RNAPII transcription machinery, such as elongation factors and the mediator complex, so their binding tends to be directly correlated with gene expression in response to heat shock.

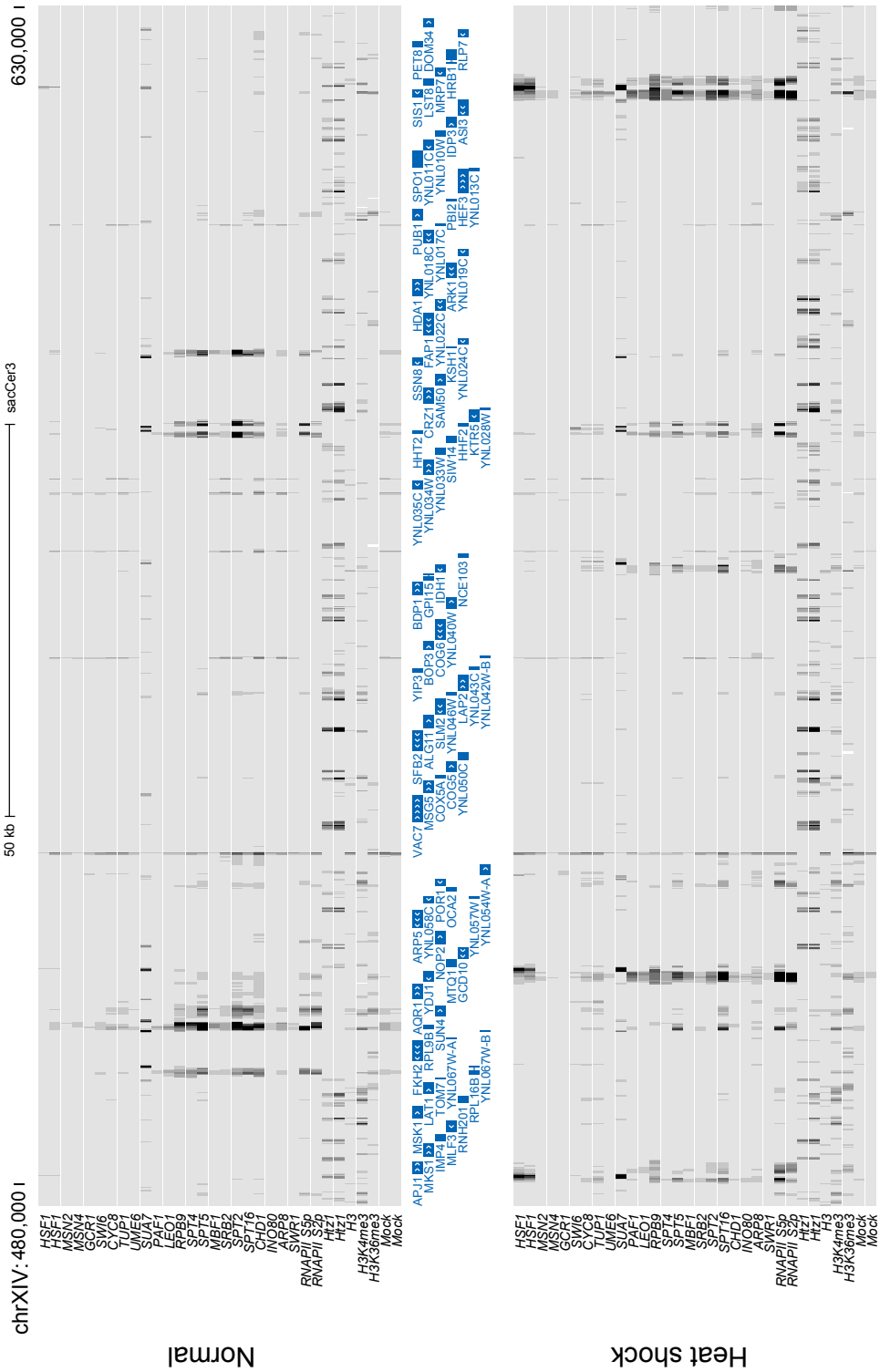


Figure 2.4: A genome browser view of the TF ChIP sequencing data. This represents

(Figure 2.4 continued.) 64 data sets including 32 independent experiments for normal (upper panel) and heat shock (bottom panel) conditions. This view includes a 150kb-genomic region from 480,000 to 630,000 bp on chromosome XIV (chrXIV). The middle panel colored in blue shows ORFs included in this region. The TF binding is depicted in black with a gradient (*e.g.*, white – no binding; gray – low binding; black – high binding of the TF).

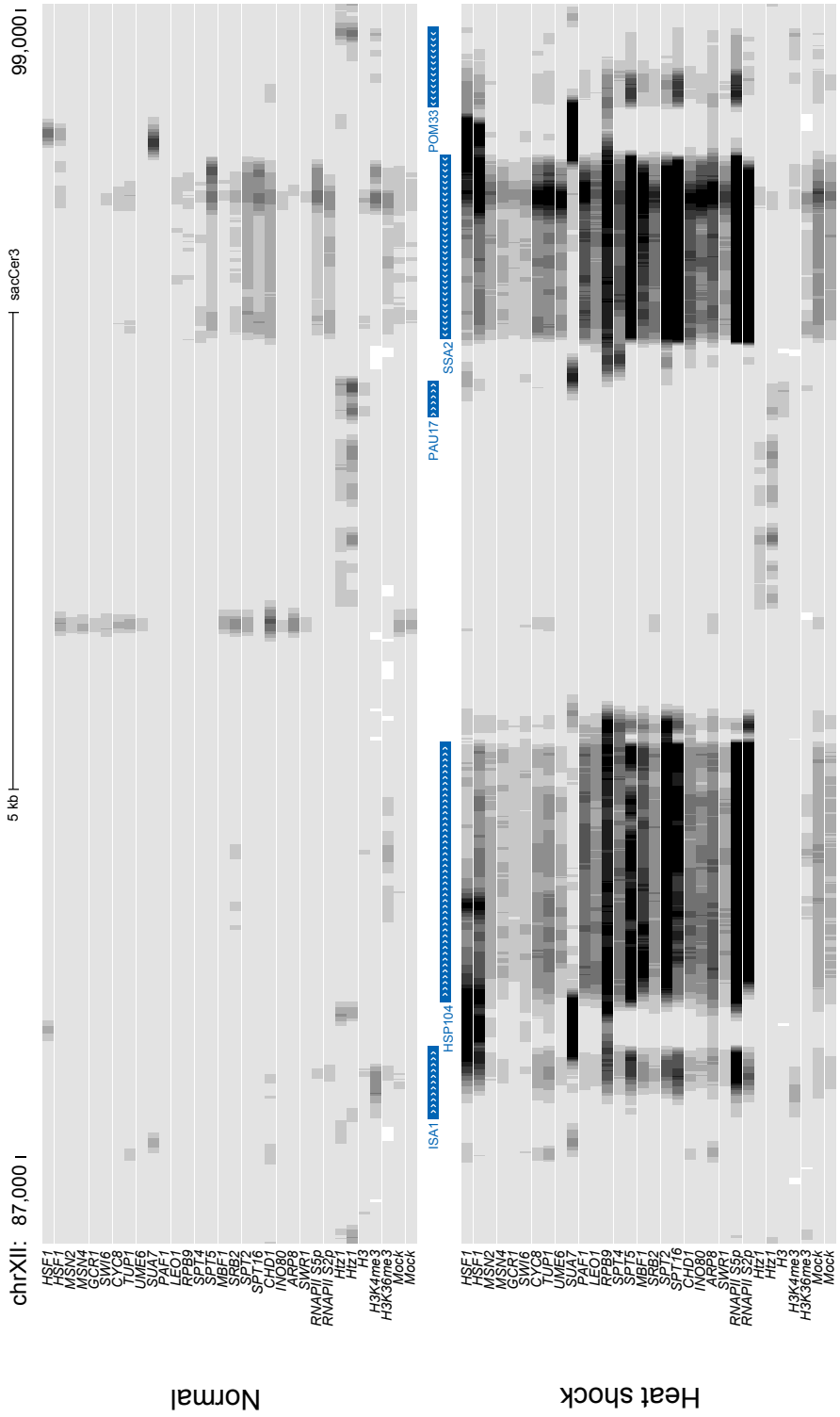


Figure 2.5: A genome browser view of the TF ChIP sequencing data. This represents 64 data sets including 32 independent experiments for normal (upper panel) and heat

(Figure 2.5 continued.) shock (bottom panel) conditions. This view includes a 12kb-genomic region from 87,000 to 99,000 bp on chrXII. The middle panel colored in blue shows ORFs included in this region. The TF binding is depicted in black with a gradient (*e.g.*, white – no binding; gray – low or background binding; black – high binding of the TF).

To understand how these TFs contribute cooperatively and distinctly to achieve control of transcription in response to heat stress, we investigated genomic loci significantly enriched by TF bindings. We conducted analysis for identifying significant binding sites (*i.e*, peaks)[i] of the TFs using MACS2 (Zhang et al., 2008). Then, we collected significant peaks by applying an adjusted p-value threshold of < 10e-10 (Figure 2.6 and Table 2.4).

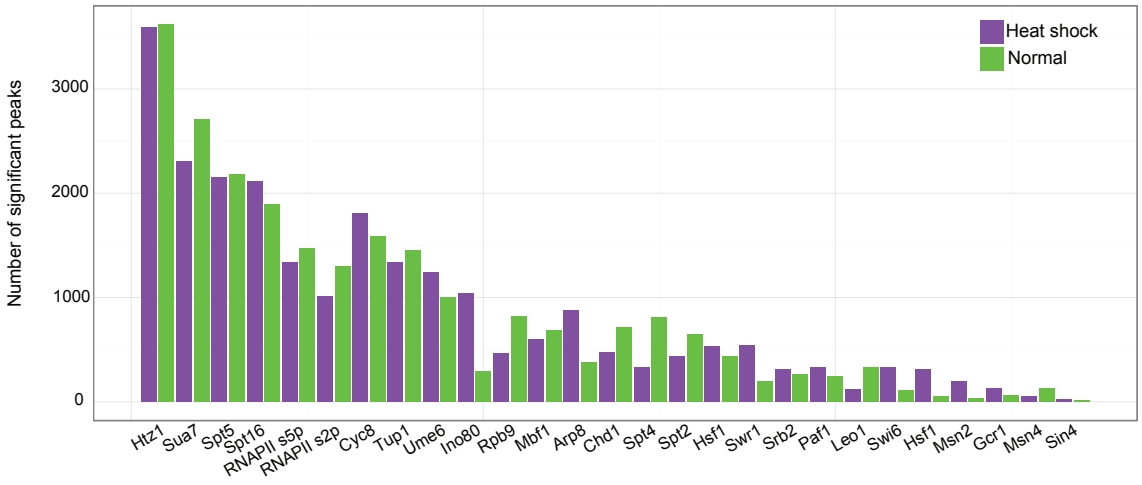


Figure 2.6: The number of significant peaks detected on the TF ChIP sequencing data obtained under normal (green) and heat shock (purple) conditions by using MACS2 peak calling (Zhang et al., 2008). The significant peaks were defined by applying an adjusted p-value threshold of < 10e-10. *Notes* – For the Hsf1 ChIP sequencing, we have two biological replicates.

[i] A significant binding site of the TF is called a "peak", and the process for identifying peaks is called "peak calling". Peak calling is a computational method for identifying genomic loci that has been enriched with aligned sequencing reads as a consequence of performing a ChIP sequencing experiment.

Table 2.4: The number of significant peaks detected on the TF ChIP sequencing data obtained under normal and heat shock conditions by using MACS2 peak calling (Zhang et al., 2008). The significant peaks were defined by applying an adjusted p-value threshold of < 10e-10.

TF	Number of significant peaks	
	Heat shock	**Normal**
Arp8	878	383
Chd1	477	714
Cyc8	1,811	1,587
Gcr1	130	61
Hsf1	844	485
Htz1	3,594	3,617
Ino80	1,045	297
Leo1	124	335
Mbf1	598	686
Msn2	195	31
Msn4	55	129
Paf1	328	249
RNAPII S2p	1,014	1,298
RNAPII S5p	1,341	1,474
Rpb9	462	823
Sin4	25	15
Spt16	2,114	1,890
Spt2	436	649
Spt4	333	815
Spt5	2,150	2,185

(Table 2.4 continued.)

TF	Number of significant peaks	
	Heat shock	**Normal**
Srb2	313	266
Sua7	2,305	2,706
Swi6	334	115
Swr1	543	199
Tup1	1,340	1,449
Ume6	1,243	1,007
Total	47,497 peaks	

To get consensus regions for TF enrichment across the all TF ChIP sequencing datasets that we generated, we combined these peaks if there is at least 1 bp overlap using Bedtools merge (Quinlan and Hall, 2010). This resulted in 3,332 genomic loci from all 47,497 significant TF peaks. These genomic loci represent the union of the significant TF binding sites for the 24 TFs and RNAPII (*i.e.*, RNAPII S2p and RNAPII S2p) across the yeast genome.

In order to see correlation between the TFs with respect to binding loci and peak intensity across the genome, we carried out the hierarchical clustering with the quantity of TF binding[i] at these loci by using Cluster3 (Eisen et al., 1998). The resulting 3,332

[i] For the quantification of TF binding, the TF ChIP sequencing reads were counted using Bedtools multicov Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

genomic loci revealed extensive changes in TF-binding in response to heat shock (Figure 2.7).

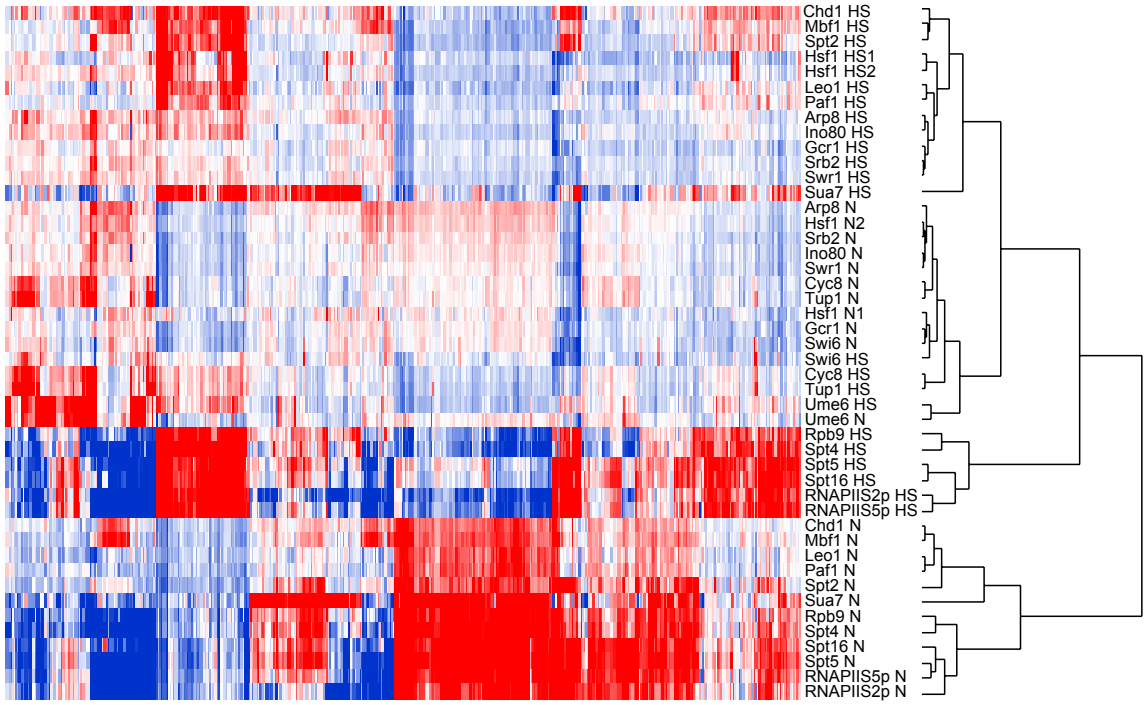


Figure 2.7: The hierarchical clustering heatmap showing differential TF-enrichment in response to heat shock. The 3,332 genomic regions, where at least one significant TF peak was detected, have been further filtered with a normalized read threshold of 0.2M in at least one TF ChIP sequencing data, resulting in 463 unique genomic loci. Each column in the heat map represents one genomic locus where binding of at least one TF is significantly high. This includes 24 TFs, RNAPIIS2p and RNAPIIS5p ChIP sequencing data generated under normal and heat shock conditions. TF bindings have been quantified by counting ChIP sequencing reads across individual genomic regions and normalizing by length and sequencing depth, and then transformed into a log2 scale. The intensity of TF binding is depicted in a blue-to-red color scheme (*e.g.*, blue – low binding; red – high binding of the TF).

2.3.3 Novel Regulation Mode of Heat Shock Transcription Factor 1 (Hsf1)

Yeast has the only one HSF, Hsf1, and its regulatory mechanism has been studied in depth. Hsf1 binds in both constitutive and stress-inducible manners to the conserved heat shock element (HSE) motif (5'-NGAAN-3') found in the promoters of its target genes (Giardina and Lis, 1995; Jakobsen and Pelham, 1988; Wiederrecht et al., 1988).

To examine genomic loci bound by Hsf1 upon heat shock, we gathered 536 and 308 significant Hsf1 peaks[i] upon heat shock from two biological replicates of Hsf1 ChIP sequencing, respectively (Figure 2.6), and this results in 621 independent loci by merging the two data sets. We discovered strong Hsf1 peaks at the promoters of well-known heat shock protein genes, such as *SSA1* and *SSA4*, as well as the promoters of genes that have never been identified as the Hsf1 target, such as *ZPR1* (Figure 2.8).

[i] An adjusted p-value threshold of < 10e-10 was applied on the output generated from MACS2 peak calling.

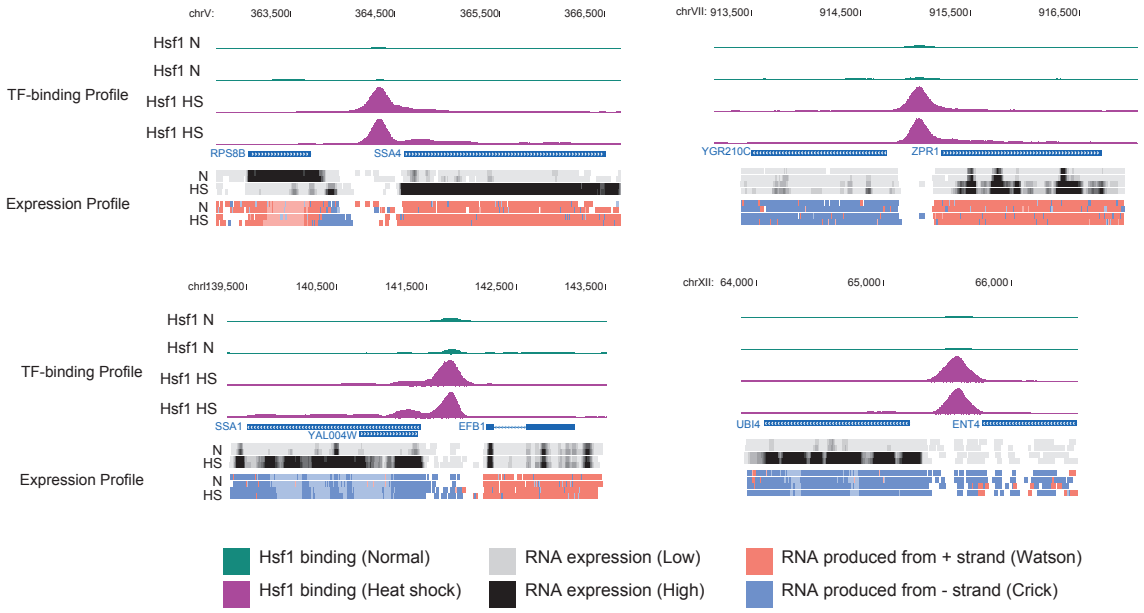


Figure 2.8: A genome browser view showing top 4 Hsf1 peaks upon heat shock. This includes the promoters of *SSA4*, *ZPR1*, *SSA1*, and *UBI4* genes. Hsf1 binding without heat shock is colored in green while Hsf1 binding upon heat shock is colored in purple. The expression level is depicted in black with a gradient (*e.g.*, white – no expression; gray – low expression; black – high expression of the transcript). The strand of RNA sequencing reads in the expression profile panel is illustrated in muted red and muted blue, Watson strand and Crick strand, respectively. The middle panel colored in blue shows ORFs included in this region. *Notes* – N: Normal, HS: Heat shock.

Unexpectedly, we found a highly reliable Hsf1 peak, which is the strongest peak on chrIX (boxed in Figure 2.9), located at an intergenic locus where no adjacent genes exist (Figure 2.10). Although there were two ORFs, approximately 1 kb upstream of the Hsf1 peak (*i.e.*, *YIR018C-A)* and 2 kb downstream of the Hsf1 peak (*i.e.*, *MUC1*), both genes did not show any expression changes upon heat stimulus (Figure 2.10), indicating that they are not Hsf1 target genes. Strikingly, two non-coding transcripts were noticeable at this locus from our RNA sequencing data (Figure 2.10). One of the non-coding

transcripts was overlapped with SUT193[i], which was identified by Xu *et al*. (2009) (Xu et al., 2009) where the authors characterized genome-wide non-coding transcripts in yeast. This short non-coding transcript was not heat-inducible, rather showed constitutively high expression. However, the other non-coding transcript that we found located at the upstream of the Hsf1 peak showed a heat-inducible expression pattern (Figure 2.10), suggesting a novel non-coding transcript regulated by Hsf1 in response to heat shock.

[i] SUT refers to stable unannotated transcripts. One type of non-coding RNAs discovered in yeast. SUTs are generally longer and more stable than cryptic unstable transcripts (CUTs), which are rapidly degraded by the exosome in cells. Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Munster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L.M. (2009). Bidirectional promoters generate pervasive transcription in yeast. Nature *457*, 1033-1037.

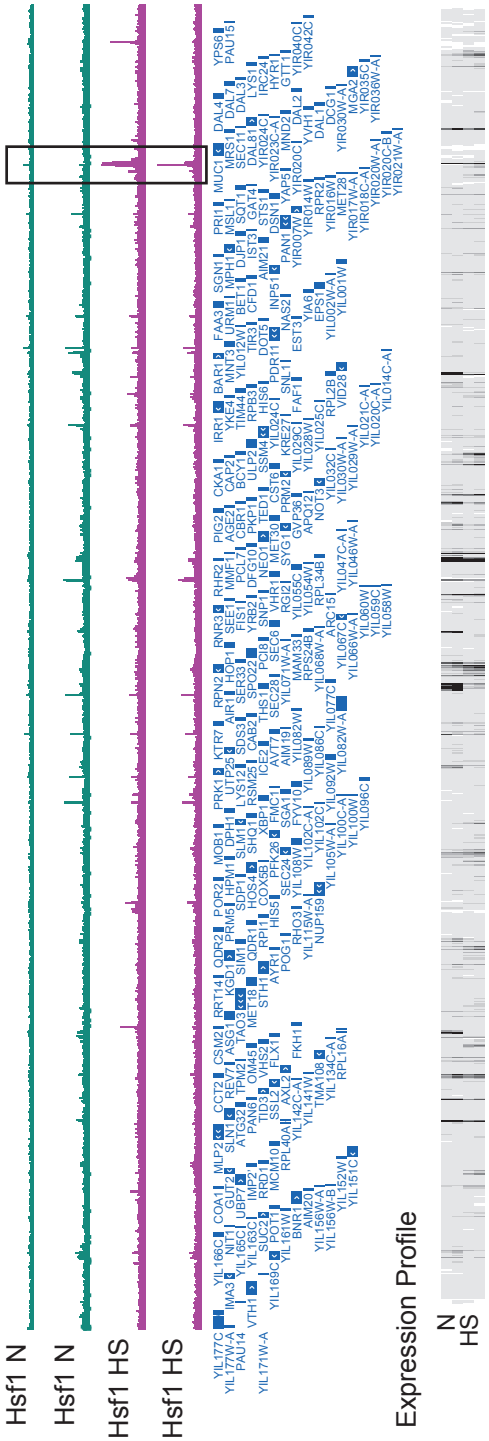


Figure 2.9: A genome browser view of the Hsf1 ChIP sequencing data and expression profile with and without heat shock in WT. This view includes the entire chrIX. Hsf1

(Figure 2.9 continued.) binding without heat shock is colored in green while Hsf1 binding upon heat shock is colored in purple. The middle panel colored in blue shows ORFs included in this region. The expression level is depicted in black with a gradient (*e.g.*, white – no expression; gray – low expression; black – high expression of the transcript). The strongest Hsf1 peak upon heat shock is marked with a black box. *Notes* – N: Normal, HS: Heat shock.

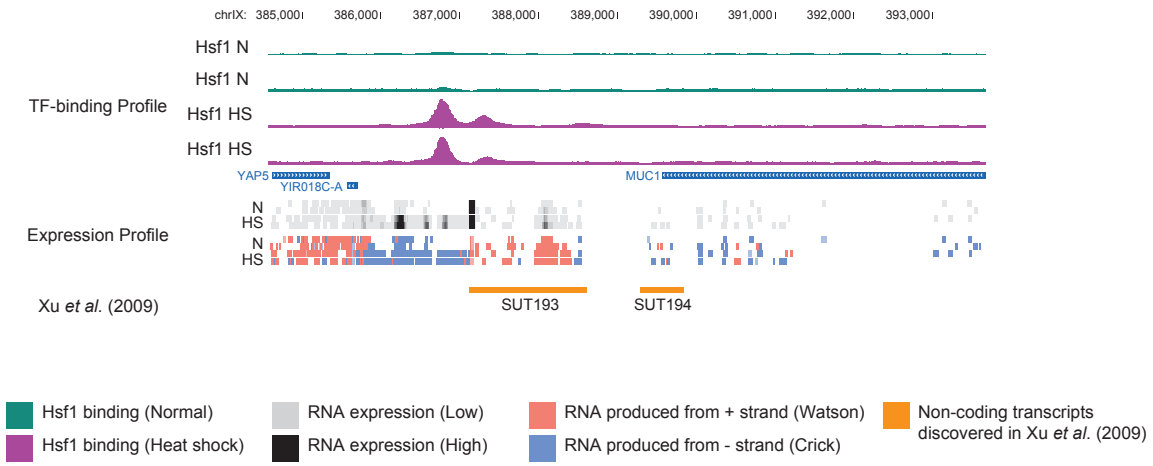


Figure 2.10: A zoomed genome browser view of the Hsf1 ChIP sequencing data and expression profile with and without heat shock in WT. This view consists of a 4-kb region on chrIX. Hsf1 binding without heat shock is colored in green while Hsf1 binding upon heat shock is colored in purple. The middle panel colored in blue shows three ORFs (*YAP5, YIR018C-A* and *MUC1*) included in this region. The expression level generated via RNA sequencing is depicted in black with a gradient (*e.g.*, white – no expression; gray – low expression; black – high expression of the transcript). The strand of RNA sequencing reads in the expression profile panel is illustrated in muted red and muted blue, Watson strand and Crick strand, respectively. Non-coding transcripts identified by Xu *et al*. 2009 (Xu et al., 2009) are illustrated in the bottom panel with orange color. *Notes* – N: Normal, HS: Heat shock, SUT: Stable Unannotated Transcript.

Additionally, we found several locations on the genome where Hsf1 seems to participate in regulating non-coding RNA transcription (Figure 2.11 and Figure 2.12). For example, *BTN2* and *SSA4* promoters are very strong Hsf1 binding sites in response to heat shock (Figure 2.11). Consistent with the high Hsf1 peaks upon heat shock, transcription of *BTN2* and *SSA4* were highly up-regulated in our RNA sequencing data (Figure 2.11), indicating these genes are Hsf1 targets in response to heat shock. Interestingly, we found heat shock-inducible non-coding RNAs upstream of these Hsf1 binding sites for both cases, which were transcribed in the anti-sense direction (Figure

2.11). This suggests that the *BTN2* and *SSA4* promoters generate bi-directional transcripts, both coding and non-coding RNAs, upon the Hsf1 binding in response to heat shock. Moreover, another strong binding of Hsf1 to the *STE20* promoter upon heat stimulus (Figure 2.12) did not affect expression of the nearby coding transcripts, *YHL008C* and *STE20*. However, we identified a heat shock-specific non-coding transcript generated at the downstream of the Hsf1 peak (Figure 2.12).

These results indicate that multiple Hsf1 peaks, which seemingly little function in gene expression, play an important role in regulating non-coding transcription, and it would be interesting to identify heat shock-induced non-coding transcripts that are regulated by Hsf1 across the yeast genome.

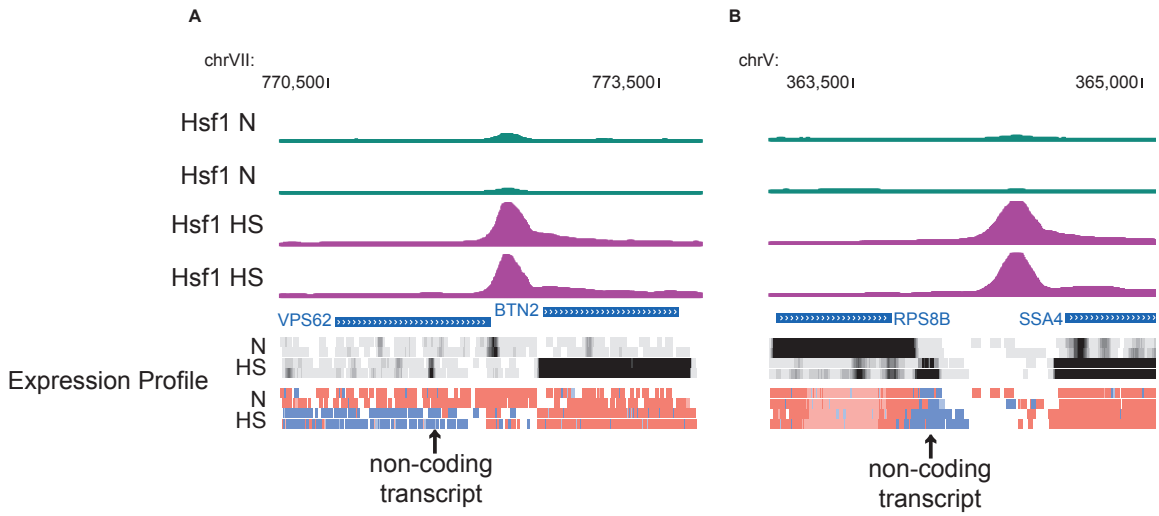


Figure 2.11: A zoomed genome browser view of the Hsf1 ChIP sequencing data and expression profile with and without heat shock in WT. (A) This view consists of a 3-kb region on chrVII including two ORFs, *VPS62* and *BTN2*, colored in blue. (B) This view consists of a 1.5-kb region on chrV including two ORFs, *RPS8B* and *SSA4*, colored in blue. Hsf1 binding without heat shock is colored in green, while Hsf1 binding upon heat shock is colored in purple. The expression profile generated via RNA sequencing is depicted in black with a gradient (*e.g.*, white – no expression; gray – low expression; black – high expression of the transcript). The strand of RNA sequencing reads in the bottom panel is illustrated in muted red and muted blue, Watson strand and Crick strand, respectively. *Notes* – N: Normal, HS: Heat shock.

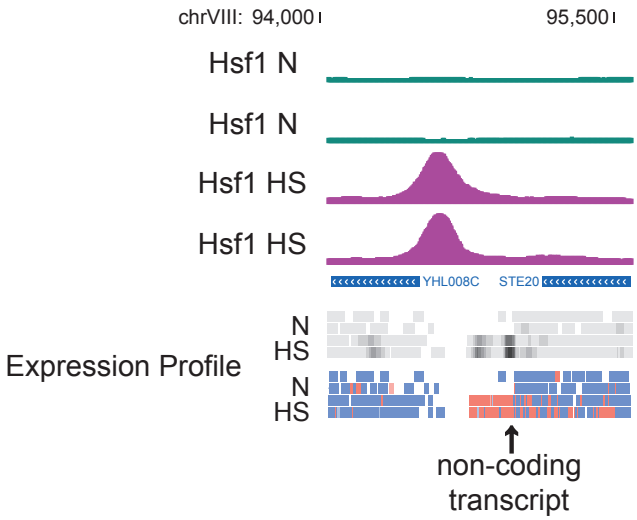


Figure 2.12: A zoomed genome browser view of the Hsf1 ChIP sequencing data and expression profile with and without heat shock in WT. This view consists of a 1.5-kb region on chrVIII including two ORFs, *YHL008C* and *STE20*, colored in blue. Hsf1 binding without heat shock is colored in green, while Hsf1 binding upon heat shock is colored in purple. The expression profile generated via RNA sequencing is depicted in black with a gradient (*e.g.*, white – no expression; gray – low expression; black – high expression of the transcript). The strand of RNA sequencing reads in the bottom panel is illustrated in muted red and muted blue, Watson strand and Crick strand, respectively. *Notes* – N: Normal, HS: Heat shock.

2.3.4 Gene Expression Changes in Transcription Factor Mutants in response to Heat Shock

If expression of a gene is changed in response to heat shock in WT, and the change is abolished in the absence of a particular TF, this indicates that the TF is responsible for inducing proper change in expression upon heat shock.

To identify genes that change expression in WT cells in response to heat shock but do not show a comparable change in cells lacking the TF, we performed gene expression profiling for the 13 TF deletion mutants and the WT with and without heat shock through microarray and RNA sequencing platforms (summarized in Table 2.2).

For quantification and comparison between microarray data sets, we used NimbleScan (Roche Nimblegen) and Bioconductor Limma (Ritchie et al., 2015; Smyth, 2004), respectively. For RNA sequencing data, we used Kallisto (Bray et al., 2016) for quantifying abundances of transcripts, and performed analysis to find genes showing differential expression using Sleuth (Harold J Pimentel, 2016).

Our analysis aimed to identify genes that differentially respond to heat shock between the TF mutant and WT. In the statistics point of view, this can be achieved to perform a hypothesis testing to determine if there is a significant interaction between two explanatory variables (*i.e.*, strain (mutant and WT), and condition (normal and heat shock)) with respect to the response variable (*i.e.*, gene expression) (Ritchie et al., 2015). Accordingly, we built the linear model considering the interaction effect for the Limma and Sleuth analyses to assess significance on the interaction effect across all genes (Figure 2.13).

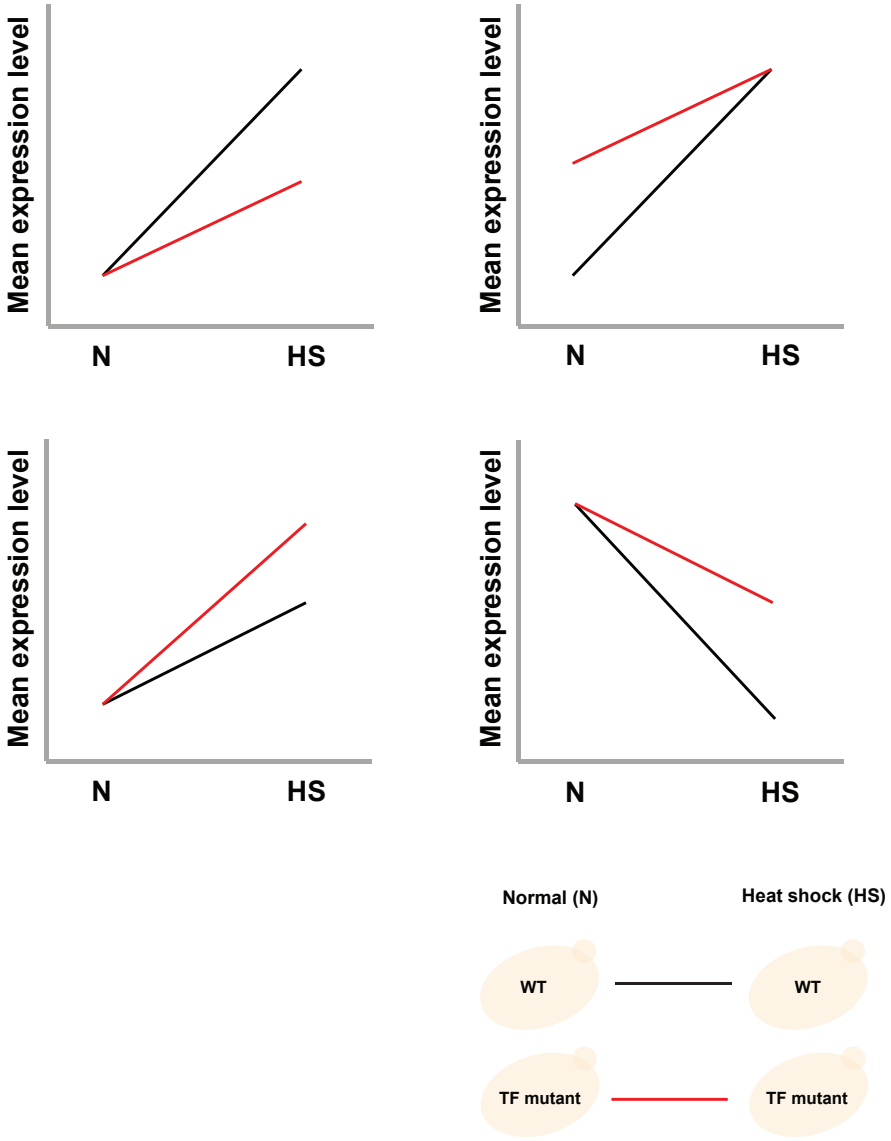


Figure 2.13: A schematic representation of analysis for identifying genes showing a significant interaction. A response variable in the linear model in this analysis is gene expression level and two explanatory variables are strains (*i.e.*, WT and TF mutant) and conditions (*i.e.*, N and HS). If a gene shows a significantly different transcriptional response upon heat shock in the TF mutant (red), shown in these four examples, compared to normal heat shock response (black), then we identified it as the differentially responded genes (DRGs). *Notes* – N: Normal, HS: Heat shock.

We applied an adjusted p-value threshold of 0.01, and consequently, identified the differentially responded genes (DRGs) to heat shock for each of the TF mutants (Table 2.5 and Figure 2.14). The representative examples of the DRG identified in the *msn2Δmsn4Δ* mutant were shown in Figure 2.15.

Table 2.5: The number of genes showing differential response to heat shock in the 13 TF mutants. The differentially responded genes (DRGs) were defined by applying an adjusted p-value threshold of < 0.01.

Strain	Technology	Adjusted p-value < 0.01 Number of DRGs
swi6Δ	RNA sequencing	1,102
tup1Δ	RNA sequencing	1,037
msn2Δmsn4Δ	RNA sequencing	556
cyc8Δ	Microarray	1,234
paf1Δ	Microarray	246
rpb9Δ	Microarray	145
sin4Δ	Microarray	51
spt4Δ	Microarray	23
arp8Δ	Microarray	22
srb2Δ	Microarray	0
gcr2Δ	Microarray	1
spt2Δ	Microarray	0
mbf1Δ	Microarray	1
Total		4,418 (2,172 unique genes)

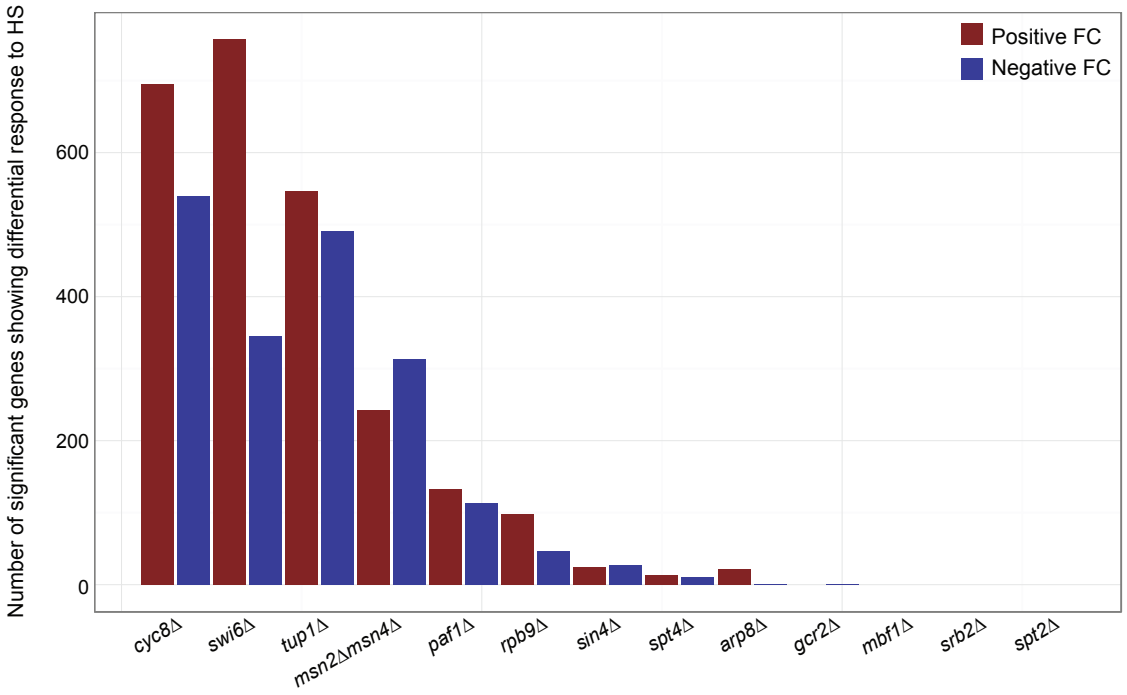


Figure 2.14: The number of genes that respond differently to heat shock in the 13 TF mutants. The differentially responded genes (DRGs) were defined by applying an adjusted p-value threshold of < 0.01. The genes showing a positive fold change (FC) are colored in red, while the genes showing a negative fold change (FC) are colored in blue. *Notes* – Positive FC: excessive expressional change in response to heat shock in the TF mutant, Negative FC: inadequate expressional change in response to heat shock in the TF mutant).

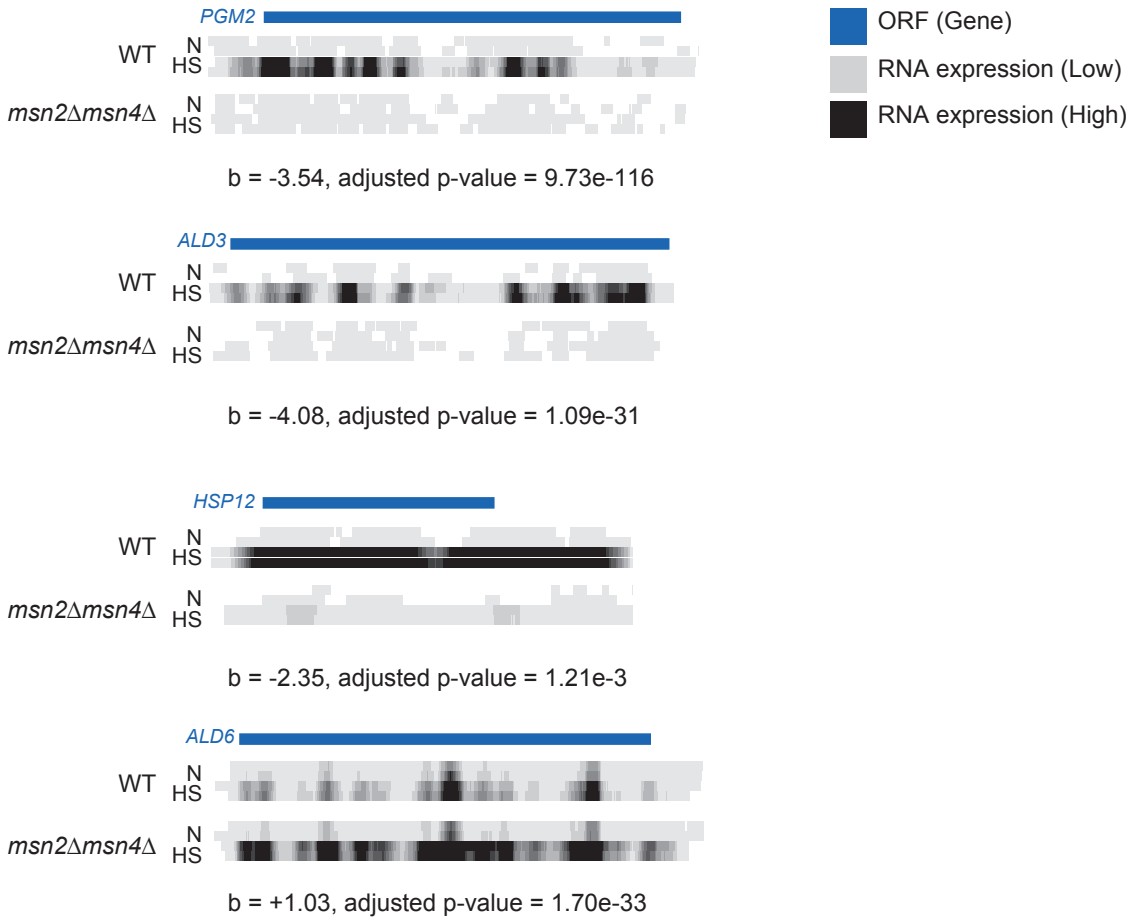


Figure 2.15: The representative examples of genes that respond differently to heat shock in the *msn2Δmsn4Δ* mutant. Four genes, *PGM2*, *ALD3*, *HSP12*, and *ALD6* are included. The gene names and ORFs are colored in blue. The expression profile generated via RNA sequencing is depicted in black with a gradient (*e.g.*, white – no expression; gray – low expression; black – high expression of the transcript). The differential analysis was carried out by using Sleuth (Harold J Pimentel, 2016), and the test outputs (*i.e.*, b – analogous to fold change – and adjusted p-value) have been shown at the bottom of each expression profile.

To elucidate interrelationship between the TFs and to assess contributions of the individual TFs systematically in order to induce transcriptional change in response to heat

shock, we conducted hierarchical clustering with the union of the DRGs by applying more stringent thresholds (*i.e.*, adjusted p-value of < 0.01 and absolute fold change > 2) (Figure 2.16). As a result, we identified 782 genes that change expression in WT cells in response to heat shock but do not show a comparable change in these TF-mutants. The resulting 782 genes would be useful for functional validation on the TF-target interactions.

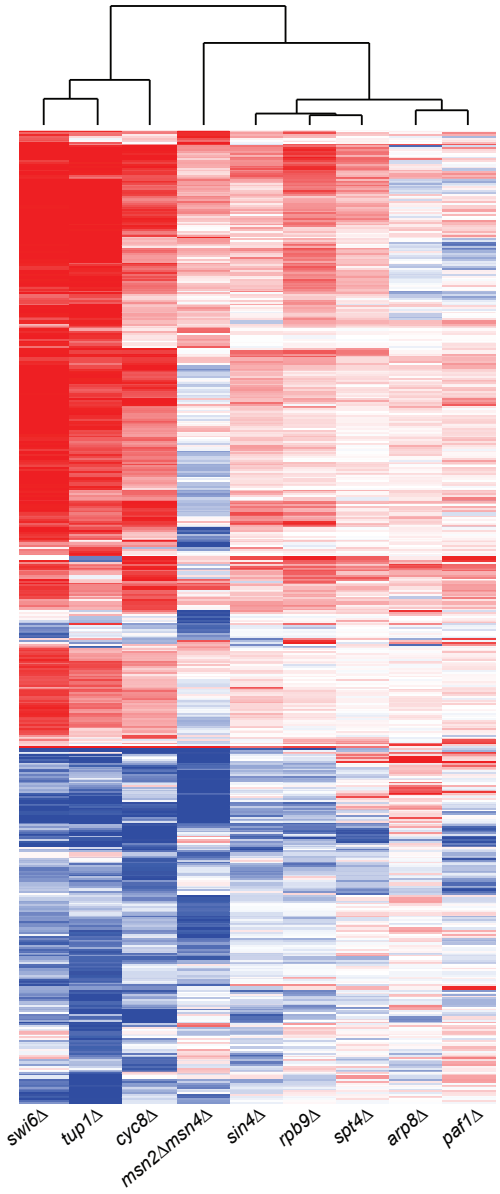


Figure 2.16: The hierarchical clustering heatmap showing differential response to heat shock in the 9 TF deletion mutants (*swi6Δ*, *tup1Δ*, *cyc8Δ*, *msn2Δmsn4Δ*, *sin4Δ*, *rpb9Δ*, *spt4Δ*, *arp8Δ*, and *paf1Δ*). Each row in the heat map represents one gene showing differential response to heat shock, when compared to the normal heat shock response observed in WT[i]. Differential response to heat shock in the TF deletion is depicted in a blue-to-red color scheme (*e.g.*, blue – inadequate expressional change in response to heat shock in the TF mutant; red – excessive expressional change in response to heat shock in the TF mutant).

[i] An adjusted p-value threshold of < 0.01 and an absolute fold change of > 2 were applied.

CHAPTER3

CHROMODOMAIN HELICASE BINDING PROTEIN 1 (*CHD1*)

3.1 Introduction

By allowing the access of DNA binding proteins to DNA through the rearrangement of nucleosomes, chromatin remodelers control diverse events occurring on the eukaryotic genome including DNA replication, DNA repair, and gene expression. The CHD (Chromodomain-Helicase-DNA binding) protein family contains a number of chromatin remodelers which have been conserved across eukaryotes. Yeast has one CHD-type remodeler, *CHD1*, and Chd1 mutants exhibit a high degree of the aberrant nucleosomal structure (Gkikopoulos et al., 2011; Park et al., 2014). As part of its fundamental role in maintaining well-positioned nucleosomes, Chd1 prevents histone replacement in the wake of RNA Polymerase II (RNAPII) and, thereby, represses cryptic transcriptional initiation over coding regions (Smolle et al., 2012).

Genome-wide studies investigating Chd1 localization have revealed that it binds at highly transcribed genes, which are marked with histone H3 tri-methylated at lysine 36 (H3K36me3) via the actions of Set2 (Park et al., 2014; Skene et al., 2014; Smolle et al., 2012). Moreover, human Chd1 specifically recognizes tri-methylated H3 lysine 4

(H3K4me3) residues produced by Set1 (Sims et al., 2005), and thus Chd1 is considered an effector of this active histone modification (Lin et al., 2011; Sims et al., 2007).

In yeast, Chd1 interacts with Set1 and Set2 (Krogan et al., 2003; Zhang et al., 2005), but does not have the specificity for directly identifying tri-methylated H3K4 (Sims et al., 2005). Furthermore, Drosophila Chd1 localization cannot be adequately explained by H3K4me3 alone (Kim et al., 2016). Despite its interactions, Chd1 binding in the yeast genome is not affected by loss of Set2 (Park et al., 2014). Given the current evidence, the relationship between Chd1 and these histone modifications, as well as the determinants of Chd1 positioning remain unclear.

Studies on a number of individual genes in yeast have implicated transcriptional elongation factors, including Rtf1 (a component of PAF1 complex), Spt4-Spt5 (DSIF complex) and Spt16-Pob3 (FACT complex), in providing Chd1 binding specificity to highly transcribed genes (Biswas et al., 2007; Simic et al., 2003). Concurringly, genome-wide Chd1 positioning shows high concordance with RNA pol II phosphorylated at serine 5 of its C-terminal domain (RNAPII S5p), which is an early elongation mark (Park et al., 2014). Yet, the effects of these elongation factors on genome-wide Chd1 localization have not been tested systematically.

In Chapter 3, we attempted to elucidate the basis for genome-wide Chd1 occupancy by examining ten candidate factors, including all components of the PAF1 complex (Paf1, Ctr9, Leo1, Cdc73, and Rtf1), two histone methyltransferases (Set1 and Set2), one component of the DSIF complex (Spt4), an essential member of the Rpd3S histone deacetylase complex (Rco1), and a histone H2 variant (Htz1) (Krogan et al.,

2003; Quan and Hartzog, 2010; Simic et al., 2003; Smolle et al., 2012). By investigating how genome-wide Chd1 occupancy is affected in deletion mutants of these factors, we uncovered that the PAF1 complex (PAF1C), a RNAPII-associated factor involved in transcription elongation, is indispensable for strong Chd1 binding on actively transcribed genes. Moreover, *SPT4*, a component of the conserved DSIF complex that regulates transcription elongation, plays a role in modulating Chd1 recruitment.

Since our two histone methyltransferase mutants, *set1Δ* and *set2Δ*, produced minimally affects on Chd1 occupancy, we generated profiles for histone H3 tri-methylation at lysine 4 and 36 (H3K4me3 and H3K36me3) in *chd1Δ* strains to establish a link between Chd1 and these histone modifications. We discovered that the loss of Chd1 causes global changes in H3K4me3 and H3K36me3 patterns throughout the yeast genome. Interestingly, the aberrant methylation patterns were predominantly observed within 1kb of transcription start sites where both methylation marks co-localize. Additionally, we detected reciprocal changes in pattern between the two marks, suggesting a possible role for Chd1 in establishing or maintaining the boundaries of these histone modifications.

Intron-containing genes were overrepresented among the genes that we identified as differentially methylated genes in the *chd1Δ* strain. This led us to analyze the effects of Chd1 on intron retention within the transcripts produced at these genes. Consequently, we discovered that intron retention was significantly lower in the absence of Chd1. This suggests that Chd1 affects RNA splicing by most likely modulating rate of RNAPII elongation.

Lastly, we attempted to investigate Chd1 function in human glioblastoma cell line by analyzing differential expression in Chd1 knockout cells that are generated via the CRISPR/Cas9 genome editing system.

3.2 Materials and Methods

3.2.1 Yeast Transformation to Create Deletion Strains and Epitope-Tagged Strains

To create the deletion strains, we replaced the ORF of interest with the cassette of a selectable marker, His3MX6 or KanMX6, by homologous recombination according to the protocol used by Geitz *et al*. (2007) (Gietz and Schiestl, 2007). Similarly, for the epitope-tagged strains, the His3MX6 cassette with either the TAP tag or the 13 MYC tag was integrated at the 3' end of the ORF of interest through homologous recombination, thereby enabling the expression of C-terminal TAP-tagged or MYC-tagged proteins. The strains generated were confirmed by PCR-based genotyping. The tagged-strains were also tested by western blot to validate that the protein of interest carried the tag correctly.

3.2.2 Data Acquisition

Table 3.1 shows the data sets that we collected for this study.

Table 3.1: Summary of experimental datasets investigated in Chapter 3. *Notes* – NA: not applicable. More details about the immunoprecipitants are described in Appendix A.

Strain	Experiment	Immunoprecipitant	Replicates #
WT *CHD1-MYC*	ChIP-Seq	Anti-Myc Beads	2
paf1Δ CHD1-MYC	ChIP-Seq	Anti-Myc Beads	1
ctr9Δ CHD1-MYC	ChIP-Seq	Anti-Myc Beads	1
leo1Δ CHD1-MYC	ChIP-Seq	Anti-Myc Beads	1
cdc73Δ CHD1-MYC	ChIP-Seq	Anti-Myc Beads	1
htz1Δ CHD1-MYC	ChIP-Seq	Anti-Myc Beads	1
rtf1Δ CHD1-MYC	ChIP-Seq	Anti-Myc Beads	1
set1Δ CHD1-MYC	ChIP-Seq	Anti-Myc Beads	1
rco1Δ CHD1-MYC	ChIP-Seq	Anti-Myc Beads	1
set2Δ CHD1-MYC	ChIP-Seq	Anti-Myc Beads	1
spt4Δ CHD1-MYC	ChIP-Seq	Anti-Myc Beads	1
WT	ChIP-Seq	Anti-K4me3 Antibody	2
WT	ChIP-Seq	Anti-K36me3 Antibody	2
chd1Δ	ChIP-Seq	Anti-K4me3 Antibody	2
chd1Δ	ChIP-Seq	Anti-K36me3 Antibody	2
WT	RNA-Seq	NA	2
chd1Δ	RNA-Seq	NA	2
WT	Microarray	NA	2
paf1Δ	Microarray	NA	2
spt4Δ	Microarray	NA	2

3.3 Results and Discussion

3.3.1 Candidate Factors Affecting Chd1 Binding Specificity in the Yeast Genome

To identify factors that give rise to the specificity of Chd1 binding, we prioritized ten candidate factors that have been reported to physically or functionally associate with Chd1, which include all components of the PAF1 complex (Paf1, Ctr9, Leo1, Cdc73, and Rtf1), two histone methyltransferases (Set1 and Set2), one component of the DSIF (Spt4), an essential member of the Rpd3S histone deacetylase complex (Rco1), and a histone H2 variant (Htz1) (Krogan et al., 2003; Quan and Hartzog, 2010; Simic et al., 2003; Smolle et al., 2012).

We created ten deletion mutants and investigated if Chd1 occupancy is affected in the deletions when compared with wild-type (WT). All deletions of PAF1C components, with the exception of *rtf1Δ*, showed significantly lower Chd1 occupancy, while *spt4Δ* revealed more Chd1 binding, when we considered top 100 genes highly bound by Chd1 and top 500 genes that are highly expressed in normal cells (Figure 3.1 and Figure 3.2).

We observed the greatest difference in Chd1 binding in *paf1Δ*, *ctr9Δ*, and *spt4Δ* (Figure 3.1); to examine how Chd1 binding was altered in these strains, we generated average Chd1 binding profiles from 1 kb upstream of the transcription start site (TSS) to 2 kb downstream of the TSS for actively transcribed genes (Figure 3.2).

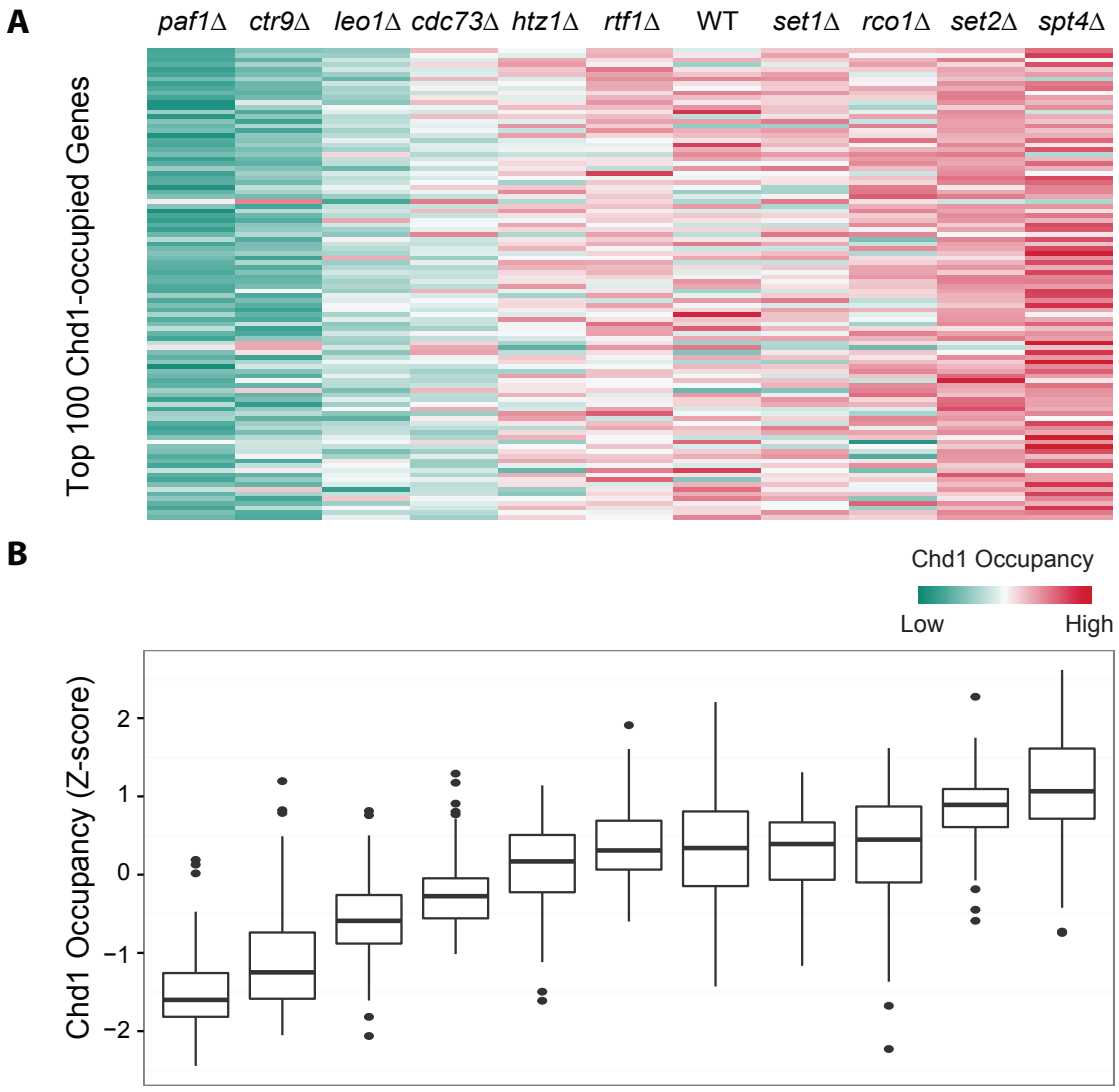


Figure 3.1: Chd1 occupancy in the ten deletion mutants and in the WT strain at genes which are highly bound by Chd1. (A) The heat map shows Chd1 occupancy in these mutants at 100 genes where Chd1 occupancy is high in WT cells. Chd1 occupancy is measured by counting Chd1 ChIP sequencing reads from the transcription start site (TSS) to transcription termination site (TTS) across the transcript and normalizing by transcript length and sequencing depth. The level of occupancy is depicted in a green-to-red color scheme (*e.g.*, green – low; red – high occupancy of Chd1) after being standardized into z-score per row (*i.e.*, gene). (B) Boxplots demonstrate the distribution of Chd1 occupancy by strain for top 100 genes with high occupancy of Chd1 in WT cells.

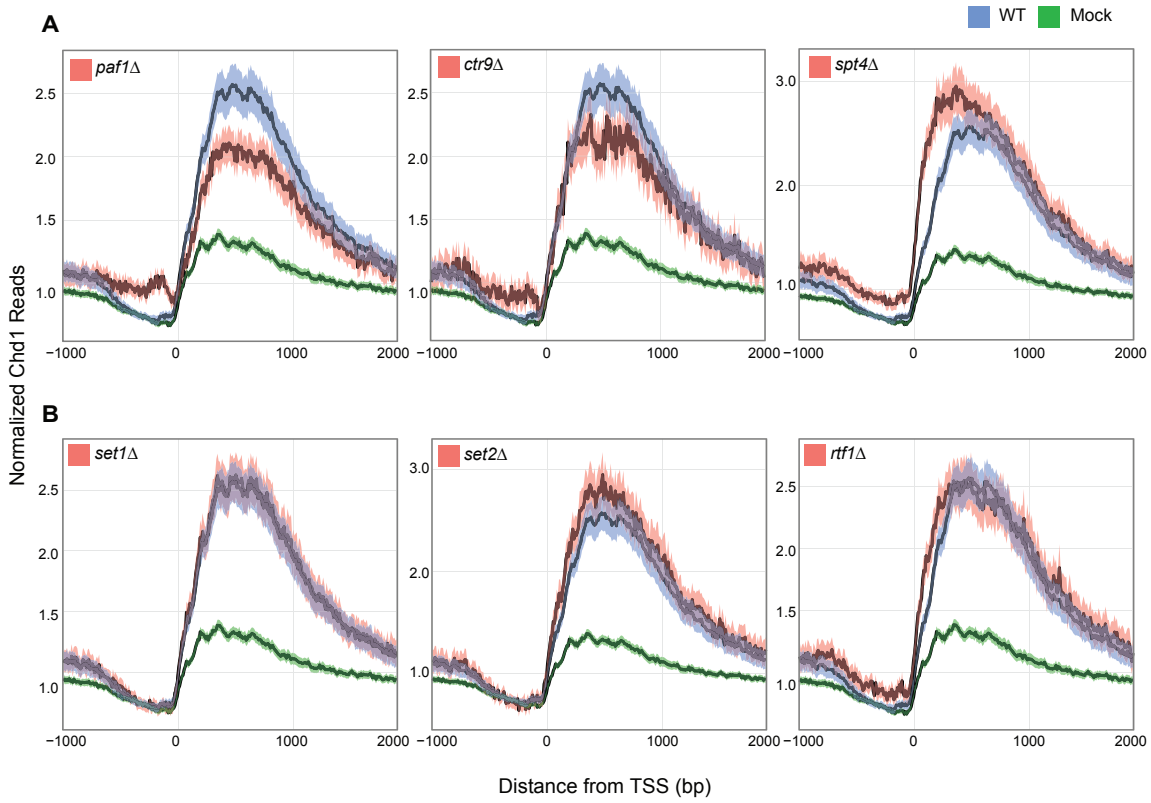


Figure 3.2: Average Chd1 occupancy profiles of selected deletion mutants from 1 kb upstream to 2 kb downstream of the transcription start sites (TSS) for highly transcribed genes. (A) Paf1, Ctr9, and Spt4 mutants. (B) Set1, Set2, and Rtf1 mutants. Chd1 occupancies measured in the mutant, WT, and mock control are colored in red, blue (positive control), and green (negative control), respectively. A line indicates the mean of the normalized Chd1 ChIP sequencing reads for the highly transcribed genes, and the shaded area around the line is a 95% confidence interval for the mean. *Notes* – The unit for normalized Chd1 reads on the y-axis is RPM (reads per million).

The *PAF1* and *CTR9* mutants revealed a considerable depletion over the gene body with a local accumulation of Chd1 at the promoter (Figure 3.2A). This suggests that PAF1C plays a role in moving Chd1 forward at highly transcribed genes. In contrast to the PAF1C mutants, *spt4Δ* showed significantly more Chd1 binding near the TSS (Figure

3.2A), indicating that Spt4 is involved in modulating Chd1 recruitment to the 5' ends of the genes.

We confirmed that *PAF1* and *SPT4* effects on Chd1 occupancy are not due to a change in gene expression in *paf1Δ* or *spt4Δ*, by investigating average Chd1 binding profiles excluding genes that are differentially expressed in either the *PAF1* or the *SPT4* mutant (Figure 3.3).

Rtf1 (a PAF1C component) has been determined to be a recruitment factor for Chd1 in yeast because of its direct interaction with Chd1 (Simic et al., 2003). Nevertheless, *rtf1Δ* showed neither a notable change in Chd1 occupancy (Figure 3.1) nor a similar pattern compared to the other members of the PAF1C (Figure 3.2B). Moreover, loss of *SET1* or *SET2*, two histone H3 methyltransferases known for their close functional relationship with Chd1, minimally affected Chd1 occupancy (Figure 3.2B).

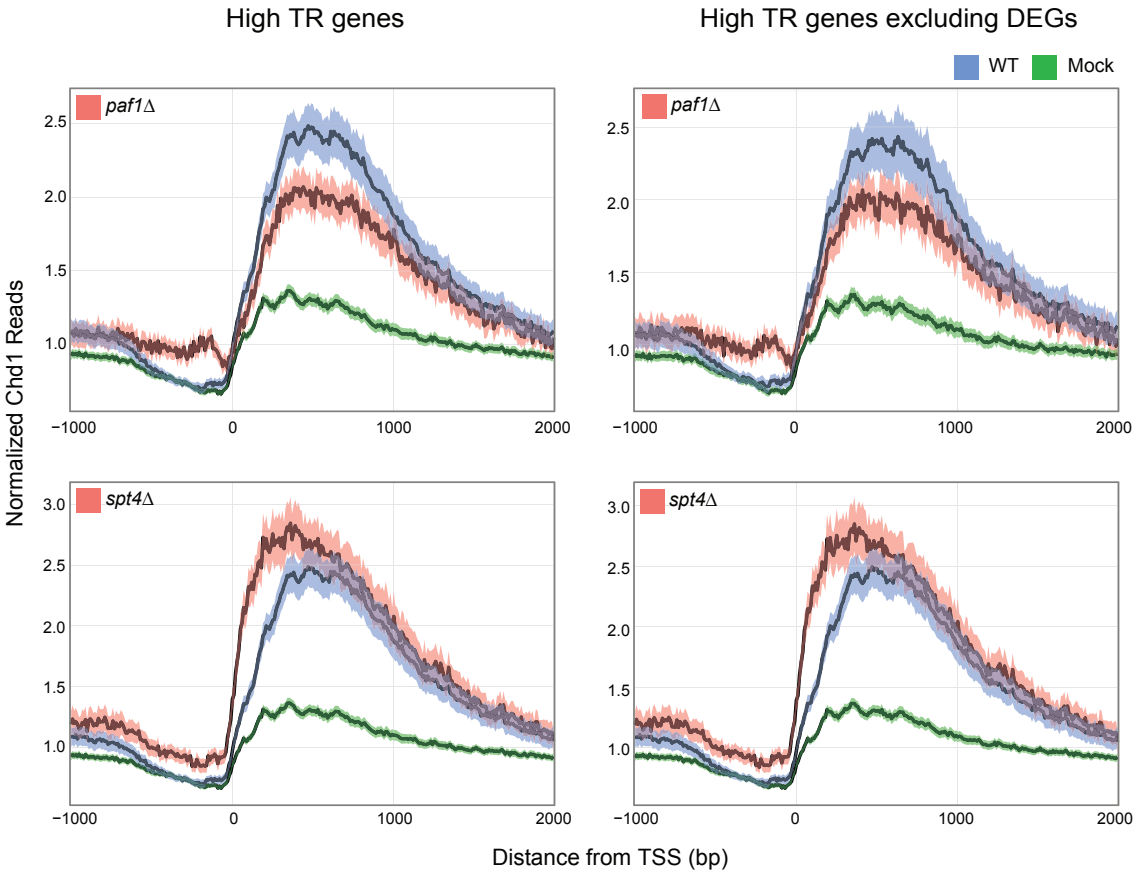


Figure 3.3: Average Chd1 occupancy profiles of the *PAF1* and *SPT4* mutants over highly transcribed genes (Left) and highly transcribed genes excluding the genes whose expression is affected by deletion of either *PAF1* or *SPT4* (Right), from 1 kb upstream to 2 kb downstream of the transcription start sites (TSS). Chd1 occupancy measured in the mutant, WT, and mock control is colored in red, blue (positive control), and green (negative control), respectively. A line indicates the mean of the normalized Chd1 reads for the highly transcribed genes, and the shaded area around the line is a 95% confidence interval for the mean. *Notes* – High TR genes: Highly transcribed genes, DEGs: Differentially expressed genes. The unit for normalized Chd1 reads on the y-axis is RPM (reads per million).

In summary, we found that all components of the PAF1C, except for Rtf1, are required for normal level and distribution of Chd1 over the highly transcribed genes (Figure 3.1 and Figure 3.2A). Spt4 prevents excessive Chd1 from being recruited to the 5' region of the genes (Figure 3.2A). Unexpectedly, none of the factors tested in this study completely abolished the specificity of Chd1 recruitment to actively transcribed regions in the yeast genome.

Chd1 is engaged over the entire unit of transcription, in highly transcribed genes, from the TSS to the TTS. Given this behavior, the factor(s) maintaining Chd1 binding to chromatin is likely to have the same binding pattern as Chd1 and be necessary for fast transcription. The most plausible factor would be RNAPII itself or a factor that moves with RNAPII, such as the PAF1C or DSIF complex (Spt4-Spt5), which was investigated in this study. Beyond these two complexes, we observed that the most similar genome-wide binding profiles to Chd1 are Spt2 and Mbf1, a DNA binding protein with HMG-like domains and a RNAPII coactivator, respectively (Figure 2.8). According to the published data on Spt2 (Nourani et al., 2006), its functional roles in transcription elongation are consistent with the known roles of Chd1. Investigating the factors with the most agreement to Chd1 binding profiles will likely shed some light on Chd1 co-factors.

3.3.2 Relationship Between *CHD1* and Histone H3 Tri-Methylation at Lysine 4 and Lysine 36

Considering the strong association between Chd1, H3K4me3 via Set1, and H3K36me3 via Set2, we decided to investigate how these two histone methylation marks are affected in the absence of Chd1. Previously, this has been tested through ChIP-on-chip experiments; the authors focused on changes in methylation at the 3' ends of genes in conjunction with Chd1 effects on histone replacement in long genes (Radman-Livaja et al., 2012). We utilized our ChIP-sequencing data to generate high-resolution profiles for H3K4me3 and H3K36me3 in *chd1Δ* and WT.

Our data illustrated that *CHD1* loss causes global changes in H3K4me3 and H3K36me3 profiles across the yeast genome (Figure 3.4, Figure 3.5, and Figure 3.6). We predominantly observed changes in the methylation pattern in *chd1Δ* within 1 kb from the TSS where both marks co-localize (Figure 3.5 and Figure 3.6).

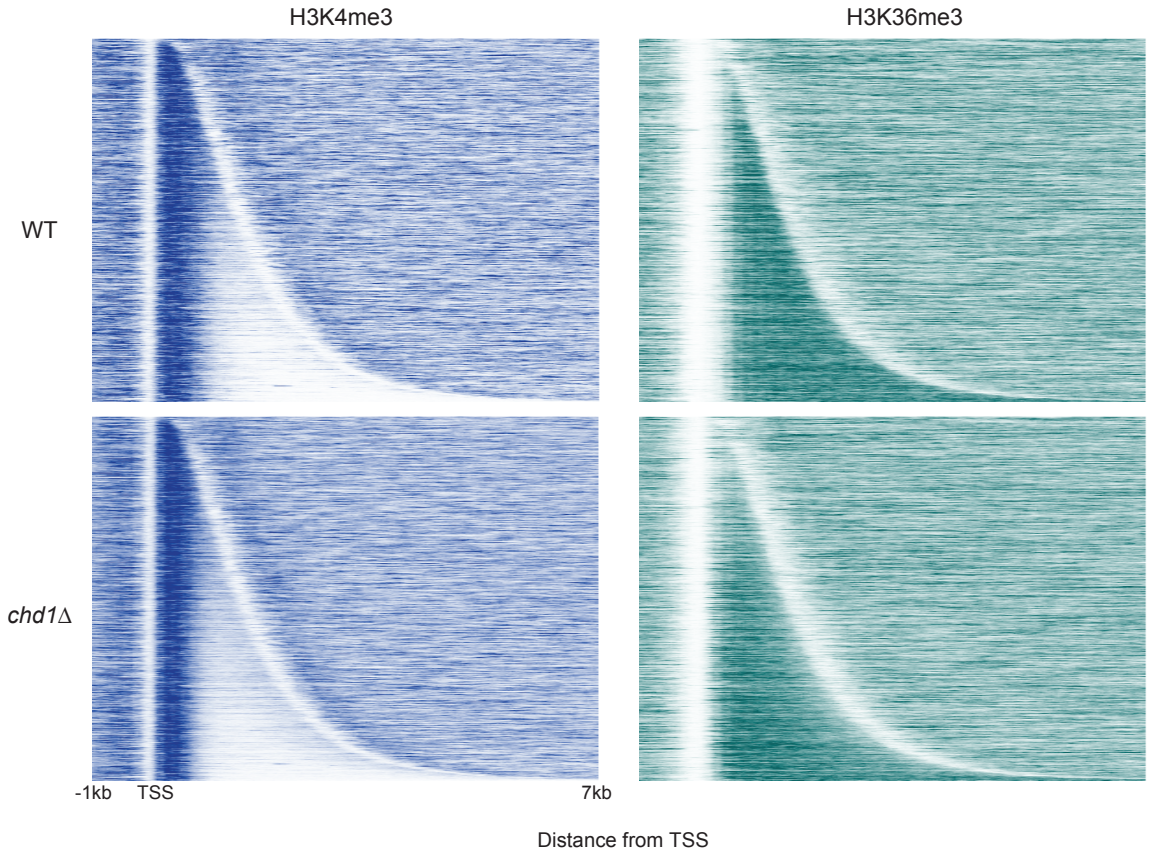


Figure 3.4: H3K4me3 and H3K36me3 patterns in WT and *chd1Δ* across all genes in the yeast genome. The methylation levels are measured by counting H3K4me3 and H3K36me3 ChIP sequencing reads from 1 kb upstream to 7 kb downstream of the TSS for all genes. The H3K4me3 and H3K36me3 signals are colored in blue and green, respectively, and the intensity of color represents the level of the methylation marks. The rows are sorted by transcript length.

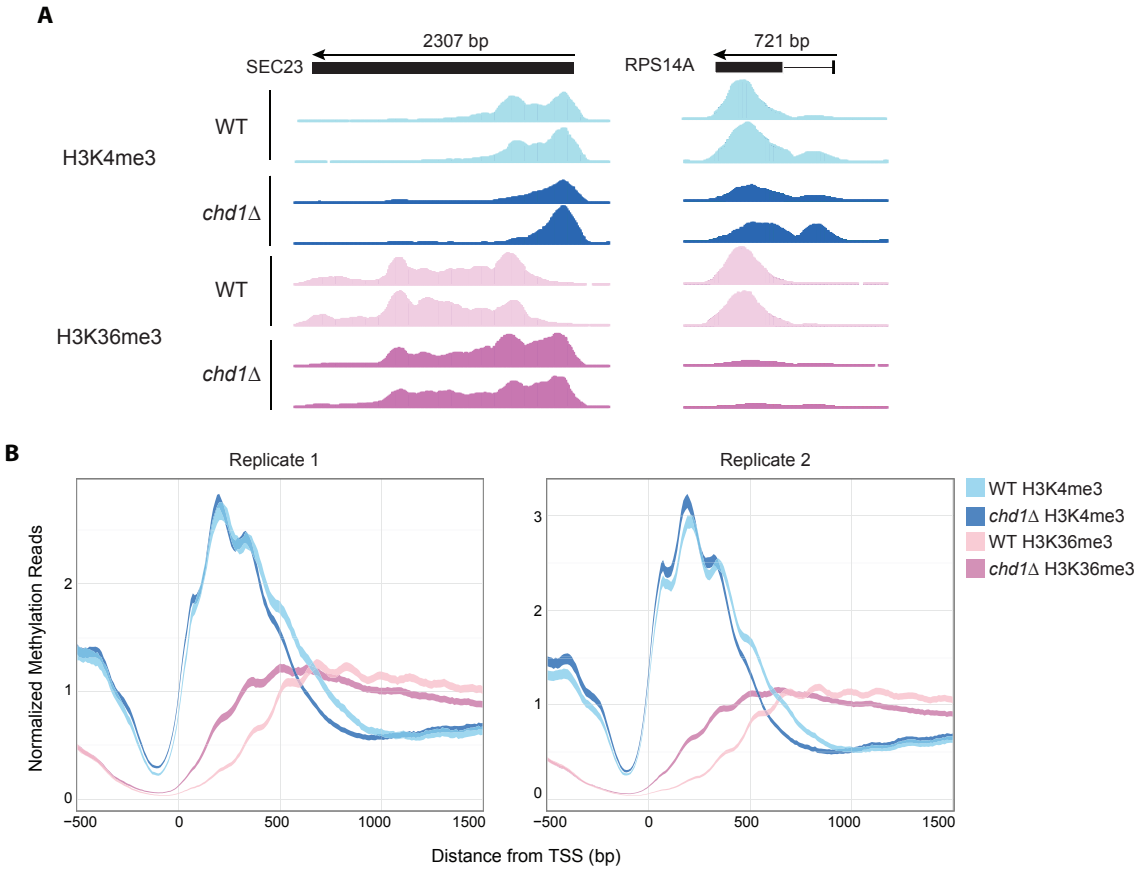


Figure 3.5: Genome-wide changes in H3K4me3 and H3K36me3 in the absence of *CHD1*. (A) Close-up views of histone tri-methylation changes across the *SEC23* and *RPS14A* genes in *chd1Δ*. (B) Average profiles of H3K4me3 and H3K36me3 for all yeast genes (for two replicate experiments) from 500 bp upstream to 1500 bp downstream of the TSS. The width of lines corresponds to a 95% confidence interval for the mean. H3K4me3 and H3K36me3 signals in WT are colored in sky blue and pink, respectively. H3K4me3 and H3K36me3 signals in *chd1Δ* are colored in dark blue and purple, respectively. *Notes* – H3K4me3: Histone H3 Lysine 4 tri-methylation, H3K36me3: Histone H3 Lysine 36 tri-methylation. *Notes* – The units for normalized reads on the y-axis are RPM (reads per million).

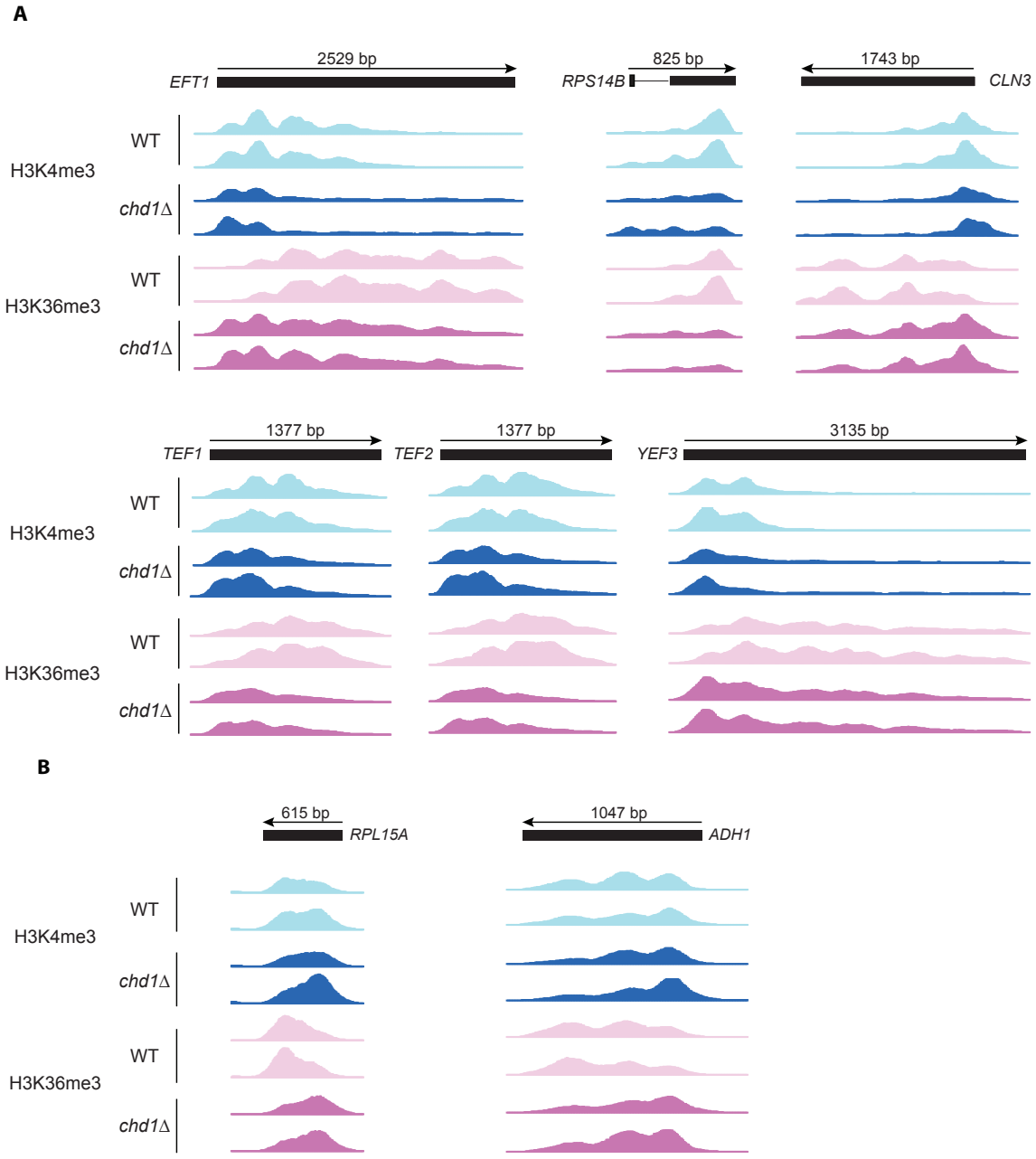


Figure 3.6: Close-up views of changes in histone tri-methylation in the *chd1Δ* strain. H3K4me3 and H3K36me3 signals in WT are colored in sky blue and pink, respectively. H3K4me3 and H3K36me3 signals in *chd1Δ* are colored in dark blue and purple, respectively. *Notes* – H3K4me3: Histone H3 Lysine 4 tri-methylation, H3K36me3: Histone H3 Lysine 36 tri-methylation.

To specifically detect the region showing differential depletion or enrichment of the histone methylation marks at genes, we used our nucleosome position data (Park et al., 2014) to bin our methylation data with respect to nucleosome position. According to the nucleosome periodicity defined in (Park et al., 2014), we determined that +1 nucleosome starts from 55 bp upstream of the TSS (*i.e.*, -55 bp) and each nucleosome spans a 165 bp region (Figure 3.7). To cover approximately 1 kb region from the TSS of all transcripts, where both H3K4me3 and H3K36me3 marks are observed, we created genomic coordinate data corresponding from +1 to +6 nucleosomes (from 55 bp upstream to 935 downstream of TSS)[i] for all transcripts (Figure 3.7). Then, we carried out analysis to identify genes showing differential H3K4me3 and H3K36me3 in *chd1Δ* at least one nucleosome position by using Bioconductor edgeR (Robinson et al., 2010).

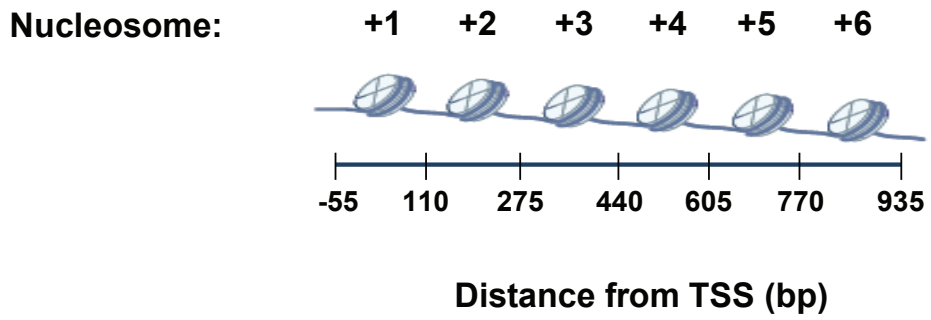


Figure 3.7: Our binning strategy to specifically detect the region showing differential depletion or enrichment of the histone methylation marks at genes.

[i] We defined the nucleosome position as follows: +1 nucleosome (-55 bp ~ +110 bp); +2 nucleosome (+110 bp ~ +275 bp); +3 nucleosome (+275 bp ~ +440 bp); +4 nucleosome (+440 bp ~ +605 bp); +5 nucleosome (+605 bp ~ +770 bp); +6 nucleosome (+770 bp ~ +935 bp). The bp in the parentheses indicates distance from transcription start sites (TSS) of transcripts.

Consequently, we identified 2,501 differentially methylated genes, using a FDR p-value less than 0.01 and greater than a 2-fold change in methylation. This set represents the union of the differentially tri-methylated genes at H3K4 and H3K36 and this number is close to half of all genes in the yeast genome (Figure 3.8).

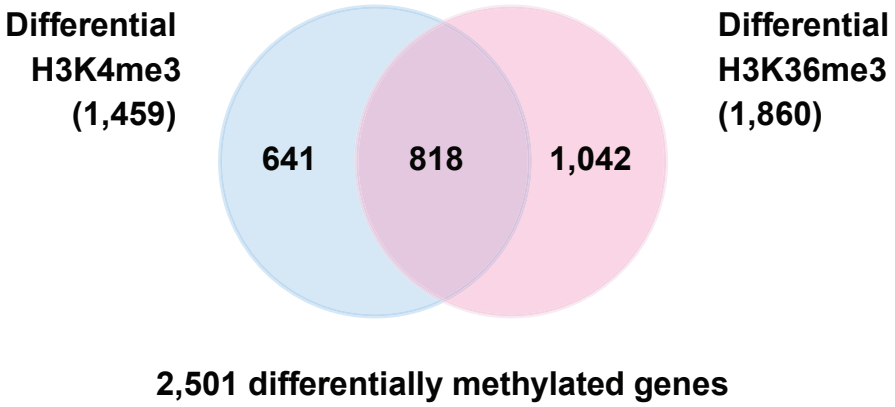


Figure 3.8: The number of genes showing differential tri-methylation at K4 and K36 of histone H3 in the absence of *CHD1*.

To elucidate methylation signal changes in the absence of Chd1, we performed hierarchical clustering with the differentially tri-methylated genes for H3K4me3 and H3K36me3 (Figure 3.9). The most striking feature of this analysis was a reciprocal change between the two methylations: if the depletion of H3K4me3 was observed at downstream nucleosomes we concurrently observed higher H3K36me3 signal at upstream nucleosomes and vice versa (Figure 3.9), indicating that Chd1 is involved in the interplay of these histone modifications.

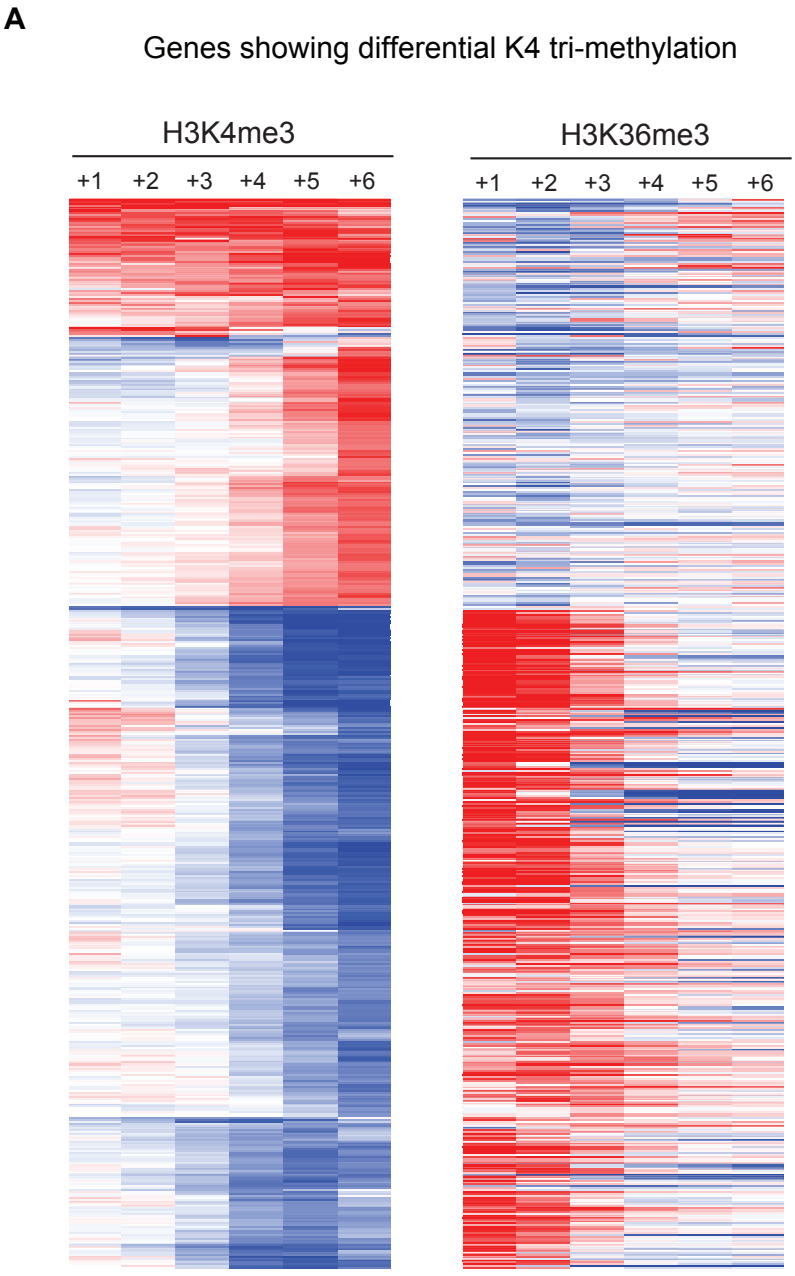


Figure 3.9: Hierarchical clustering of differential K4 and K36 tri-methylation upon loss of *CHD1*. (A) Genes showing differential K4 tri-methylation in *chd1Δ*. Hierarchical clustering was performed with the H3K4me3 data and the H3K36me3 data was incorporated to see the relationship between these two methylation marks.

(Figure 3.9 continued.)

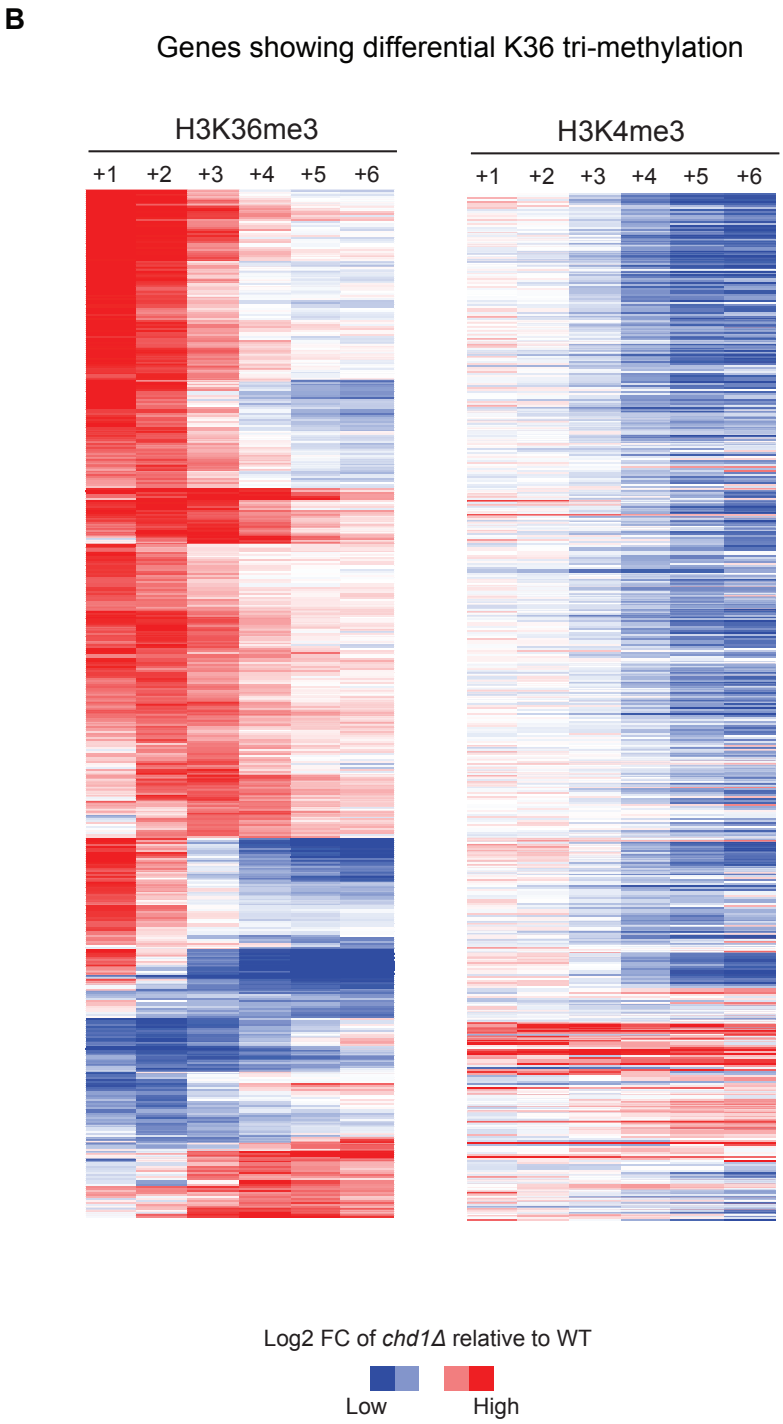


Figure 3.9: Hierarchical clustering of differential K4 and K36 tri-methylation upon loss of *CHD1*. (B) Genes showing differential K36 tri-methylation in *chd1Δ*.

(Figure 3.9 continued.) Hierarchical clustering was conducted with the H3K36me3 data and the fold-change of *chd1Δ* relative to WT in H3K4me3 was added next to it. Fold changes have been transformed into a log2 scale and depicted in a blue-to-red color scheme (*e.g.*, blue – negative fold change: less tri-methylation in *chd1Δ*, white – no fold difference between *chd1Δ* and WT, red – positive fold change: more tri-methylation in *chd1Δ*). *Notes* – +1, +2, +3, +4, +5, +6 refer to nucleosome positions and each nucleosome position is defined as a 165-bp region from 55 bp upstream to 935 bp downstream of the TSS.

In summary, we consolidated the relationship between Chd1 and two histone H3 modifications, H3K4me3 and H3K36me3, through high resolution sequencing and a binning-based strategy to detect differential methylation in the absence of Chd1. Loss of *CHD1* led to aberrant methylation patterns in approximately half of the yeast genome (Figure 3.8). These methylation changes were predominantly observed within 1kb of the transcription start sites of the genes, where both methyl marks should be present (Figure 3.4, Figure 3.5, and Figure 3.6). The changes in H3K4me3 and H3K36me3 were reciprocal between relatively upstream (+1, +2, and +3) and downstream nucleosomes (+4, +5, and +6) presented in the 1kb window (Figure 3.9). The disrupted placement of these methylations due to loss of *CHD1* was obvious and widespread across the yeast genome (Figure 3.4, Figure 3.5, and Figure 3.6). These results suggest a possible role for Chd1 in dictating distinct domains for H3K4 and K36 tri-methylations. Nevertheless, its impact on gene expression was marginal (Figure 3.10), indicating that the alteration in histone modifications little influences the RNA abundance transcribed from the coding sequence under steady-state conditions.

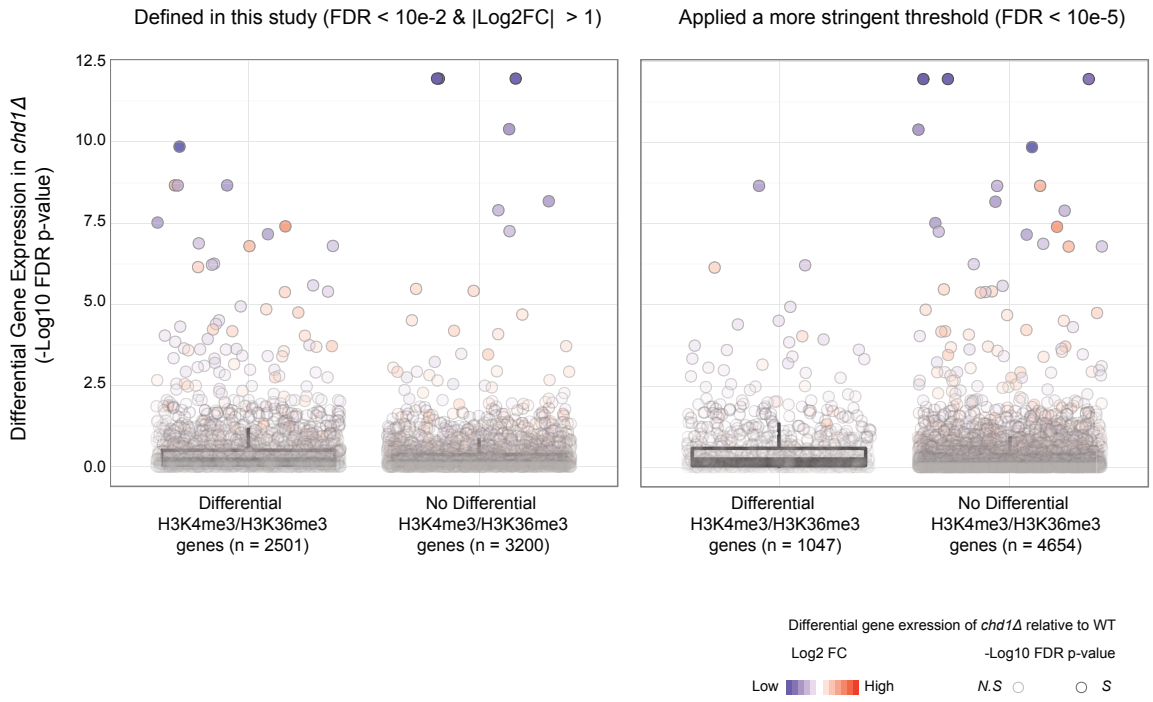


Figure 3.10: Differential expression analysis of genes in *chd1Δ* separated by whether or not the gene displays differential tri-methylation marks. (Left) The two gene categories on the x-axis (*i.e.*, differential H3K4me3/H3K36me3 genes, and no differential H3K4me3/H3K36me3 genes) are defined by a FDR threshold of < 10e-2 and an absolute log2 FC > 1. (Right) The two gene categories on the x-axis are defined by an FDR threshold of < 10e-5. Each point represents one gene with two kinds of information: the fold change and the FDR p-value, which are obtained from a differential gene expression analysis comparing the WT and *chd1Δ* RNA sequencing data by using Bioconductor edgeR (Robinson et al., 2010). The gene expression fold change of *chd1Δ* relative to WT is depicted in a blue-to-red color scheme (*e.g.*, blue – negative fold change: repressed in *chd1Δ*, white – no fold difference between *chd1Δ* and WT, red – positive fold change: activated in *chd1Δ*). The transparency of the point reflects the significance level (*e.g.*, faint – Not significant (*N.S*), sharp – Significant (*S*)).

Misregulation of histone modifications is known to lead a variety of human diseases, underscoring the significance of understanding how they are regulated on a genome-wide scale (Shilatifard, 2012; Wagner and Carpenter, 2012). Accordingly, many studies have been conducted to define the normal organization of histone modifications

with regard to position and quantity. For example, Set1 deposits a gradient pattern of methylation at H3K4 residues such as tri-, di-, and mono-methylations, from the promoter to the downstream direction of genes (Kim and Buratowski, 2009; Shilatifard, 2012; Xiao et al., 2007). Set2 can mono-, di-, and tri-methylate H3K36 on bodies of genes (Wagner and Carpenter, 2012). The K4 and K36 methylation tend to be inversely related to each other in terms of the location on genes, since as the level of H3K4 methylation diminishes, H3K36 methylation increases along genes (Narlikar et al., 2013).

One study described the well-defined tri- and di-methylated H3K4 loci by using the term "two distinct chromatin zones" (Kim and Buratowski, 2009). The authors further distinguished these two zones via the recruitment of different histone modifying enzymes (Kim and Buratowski, 2009), and the presence of overlapping non-coding transcript (Kim et al., 2012); defining disparate chromatin structures as a mechanism for fine-tuning transcriptional induction or repression.

Furthermore, a separate study revealed that Set1-mediated methylation at H3K4 determines which chromatin-remodeling factor is recruited to cause changes in chromatin structure for the condition-specific expression (Nadal-Ribelles et al., 2015).

We found that *CHD1* loss disrupted H3K4me3 and H3K36me3 domains (Figure 3.11). Considering the importance for establishing distinct chromatin domain with proper epigenetic marks, Chd1 would have high potential to affect diverse chromatin events, as seen in these previous studies.

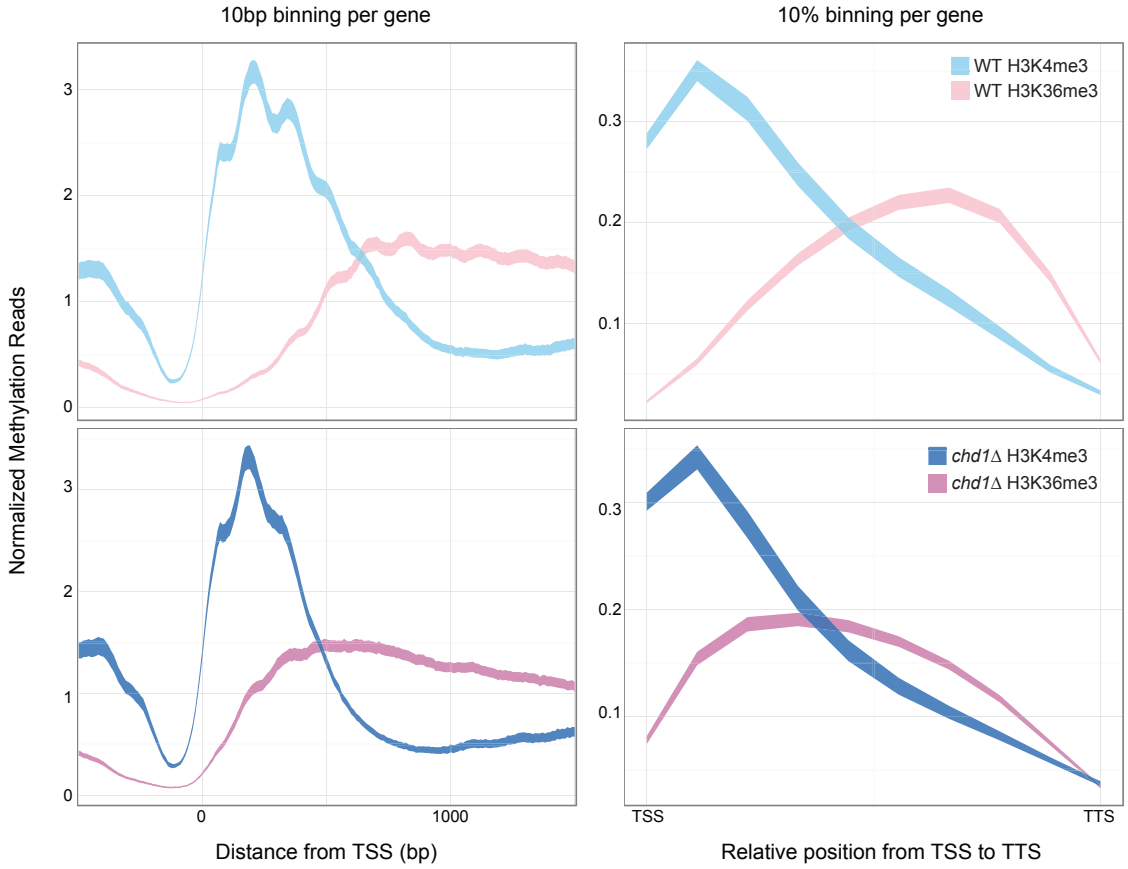


Figure 3.11: Average profiles of H3K4me3 and H3K36me3 for genes showing differential K4 and K36 tri-methylation in the absence of *CHD1*. H3K4me3 and H3K36me3 signals in WT are colored in sky blue and pink, respectively. H3K4me3 and H3K36me3 signals in *chd1Δ* are colored in dark blue and purple, respectively. (Left) Average profiles are generated by counting the methylation ChIP sequencing reads per 10-bp bin from 1.5 kb upstream to 1.5 kb downstream of TSS. (Right) Average profiles are constructed by counting the methylation ChIP sequencing reads per bin representing 10% of the transcript from TSS to TTS (transcription termination site). The width of lines corresponds to the 95% confidence interval of the mean. *Notes* – RPM: Reads per million, the unit for normalized methylation reads on the y-axis.

Additionally, we note that H3K4me3 and H3K36me3 changes in the absence of *CHD1* have been tested through ChIP-on-chip experiments in previous studies (Radman-Livaja et al., 2012; Smolle et al., 2012). One common claim was that H3K36me3 signal

shifted upstream of the genes in *chd1Δ*. Changes in H3K4me3 were underestimated due to their overall low significance. However, we found a highly significant effect at a large number of genes that show differential tri-methylation signal at K4 and K36 of histone H3 in the absence of *CHD1*. Our ability to detect this novel signature hinged on our use of ChIP-sequencing, which has a superior resolution to ChIP-chip experiments, as well as change in signal depending on the genomic loci of the genes affected. When we considered the changes between *chd1Δ* and WT by summarizing the ChIP-sequencing reads only for the gene level (*i.e.*, without binning), the number of genes that are differentially marked with these methylations underwent an 85% reduction. Representative false-negative examples (*RPL15A* and *ADH1*) are shown in Figure 3.6B. This strongly suggests the importance of analyzing histone modification data by employing a methodology to detect local signal changes such as binning (Figure 3.7) or sliding window strategies (Lun and Smyth, 2016; Schweikert et al., 2013).

3.3.3 Commonalities of Differentially Methylated Genes in the Absence of *CHD1*

To elucidate Chd1 function through the differentially tri-methylated genes in the absence of *CHD1* that we identified, we examined several characteristics of these genes. First, we examined the significance of methylation change sensitivity with *CHD1* loss as a function of transcription level and transcript length (Figure 3.12A and Figure 3.12B, respectively).

We identified a positive correlation between H3K4me3 and transcription level, specifically at +4, +5, and +6 nucleosomes (Figure 3.12A). The overall trend between the alteration in H3K36me3 upon *CHD1* loss and gene expression is also positive (Figure 3.12A), indicating that actively transcribed genes tend to show a high degree of methylation changes in the absence of *CHD1*. One striking feature was increased transcription of the genes resulted in depletion of H3K4me3 and H3K36me3 in *chd1Δ* (Figure 3.12A). However, there was no correlation between transcript length and histone methylation changes upon *CHD1* deletion (Figure 3.12B). Rather, we observed a strong bias for H3K36me3 depletion at short genes, suggesting a specific group of genes undergoes these distinct changes in methylation due to *CHD1* loss.

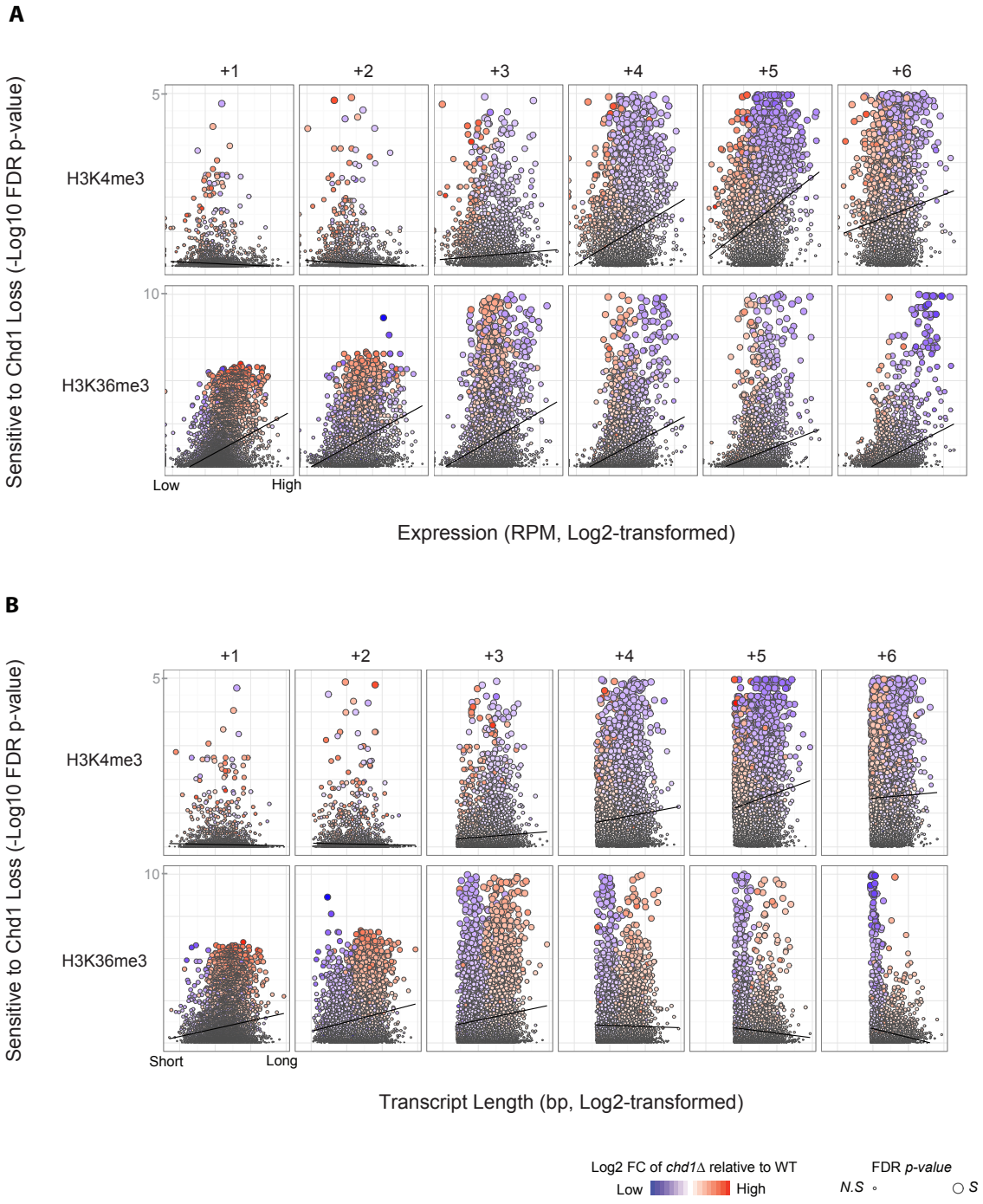


Figure 3.12: Sensitivity of H3K4me3 and H3K36me3 to *CHD1* loss with respect to transcription level (A) and transcript length (B). The sensitivity of methylation levels to *CHD1* loss has been quantified using a statistical significance level calculated based on

(Figure 3.12 continued.) changes in tri-methylation due to the *CHD1* loss. Expression values were obtained by counting RNA sequencing reads per gene in WT cells, which were then transformed into a log2 scale. Transcript lengths are the distance between TSS and TTS for individual genes, and have been transformed into a log2 scale. Fold changes of *chd1Δ* relative to WT for H3K4me3 and H3K36me3 are depicted in a blue-to-red color scheme (*e.g.*, blue – negative fold change: less tri-methylation in *chd1Δ*; white – no fold difference between *chd1Δ* and WT; red – positive fold change: more tri-methylation in *chd1Δ*). The size of the point reflects the significance level (*e.g.*, small – Not significant (*N.S*); large – Significant (*S*)). *Notes* – +1, +2, +3, +4, +5, +6 refer to nucleosome positions and each nucleosome position is defined as a 165-bp region from 55 bp upstream to 935 bp downstream of the TSS.

3.3.4 Intron Retention and Aberrant Histone Methylation Patterns in Ribosomal Protein Genes upon *Chd1* Loss

We identified the genes in the above group to be ribosomal protein (RP) genes, characterized by a high level of transcription and relatively short transcript length (Figure 3.10). We examined the aberrant histone methylation patterns on the RP genes in *chd1Δ* (Figure 3.13), and found that all 137 RP genes, except for five RP genes failing to pass the threshold (*RPL15B*, *RPS28A*, *RPS28B*, *RPL41A*, and *RPL41B*), exhibited a dramatic reduction in the tri-methylated K4 and K36 levels with the loss of *CHD1*.

The RP genes are distinctly identifiable as they are highly expressed and contain an intron; 101 out of 137 RP genes (~ 75%) have one or more intron(s). Most introns in RP genes are positioned towards the 5' end of the gene, and we observed that signal from H3K4me3 and H3K36me3 was significantly higher in these regions (Figure 3.13 and Figure 3.14). This reminded us of our previous observation that accumulation of RNAPII S5p at intron-exon junctions within highly transcribed genes in *chd1Δ* (Park et al., 2014). Moreover, we have observed that Chd1 is recruited at introns within highly transcribed genes (Figure 3.15); thus, we hypothesized that Chd1 affects RNA splicing at highly transcribed genes.

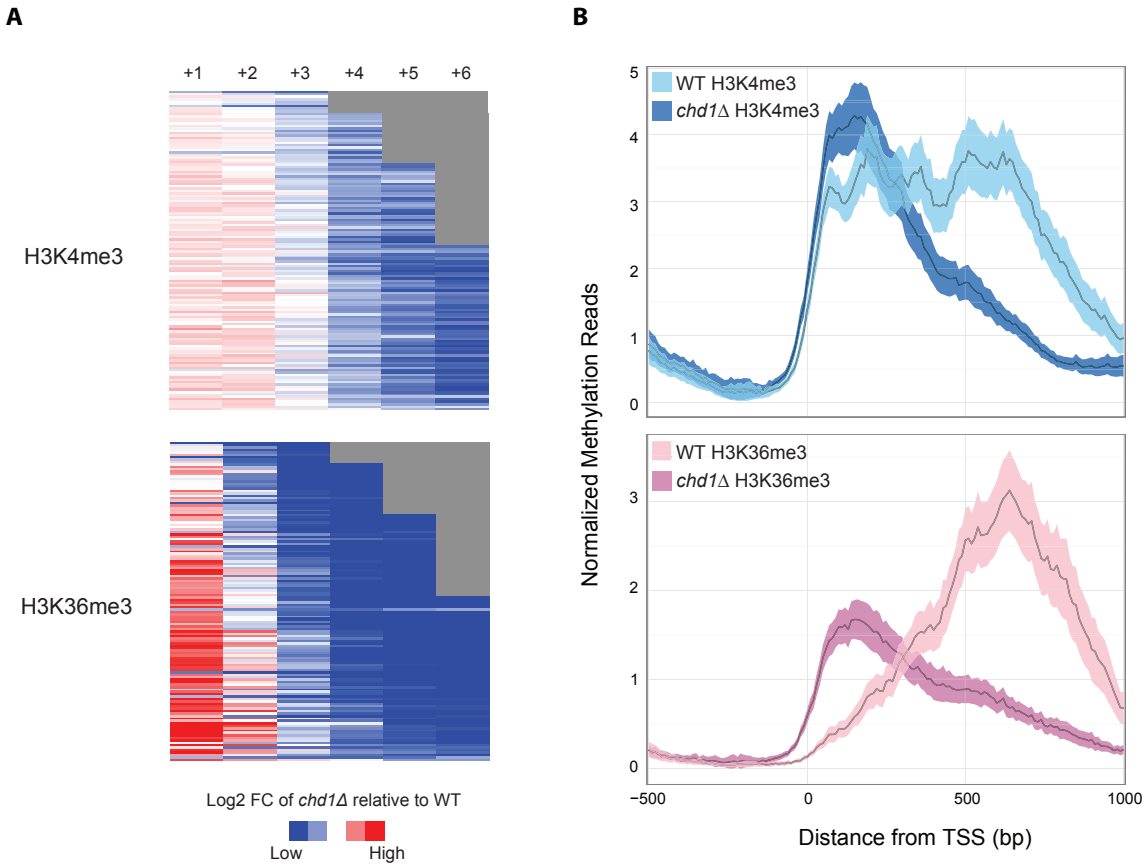


Figure 3.13: Aberrant H3K4me3 and H3K36me3 at ribosomal protein (RP) genes in the absence of *CHD1*. (A) Heatmaps showing differential H3K4me3 (top) and H3K36me3 (bottom) in *chd1Δ* relative to WT for RP genes using fold changes. The fold changes are in a log2 scale and depicted in a blue-to-red color scheme (*e.g.*, blue – negative fold change: less tri-methylation in *chd1Δ*; white – no fold difference between *chd1Δ* and WT; red – positive fold change: more tri-methylation in *chd1Δ*; grey – missing data). Each column (*e.g.*, +1, +2, +3, +4, +5, +6) represents nucleosome positions, and the rows are sorted by transcript length. (B) Average profiles of H3K4me3 (top) and H3K36me3 (bottom) for RP genes in *chd1Δ* and WT. H3K4me3 and H3K36me3 signals in WT are colored in sky blue and pink, respectively. H3K4me3 and H3K36me3 signals in *chd1Δ* are colored in dark blue and purple, respectively. The middle line indicates the mean of the normalized methylation reads for RP genes, and the shaded area around the line is a 95% confidence interval for the mean. The unit of the y-axis is RPM.

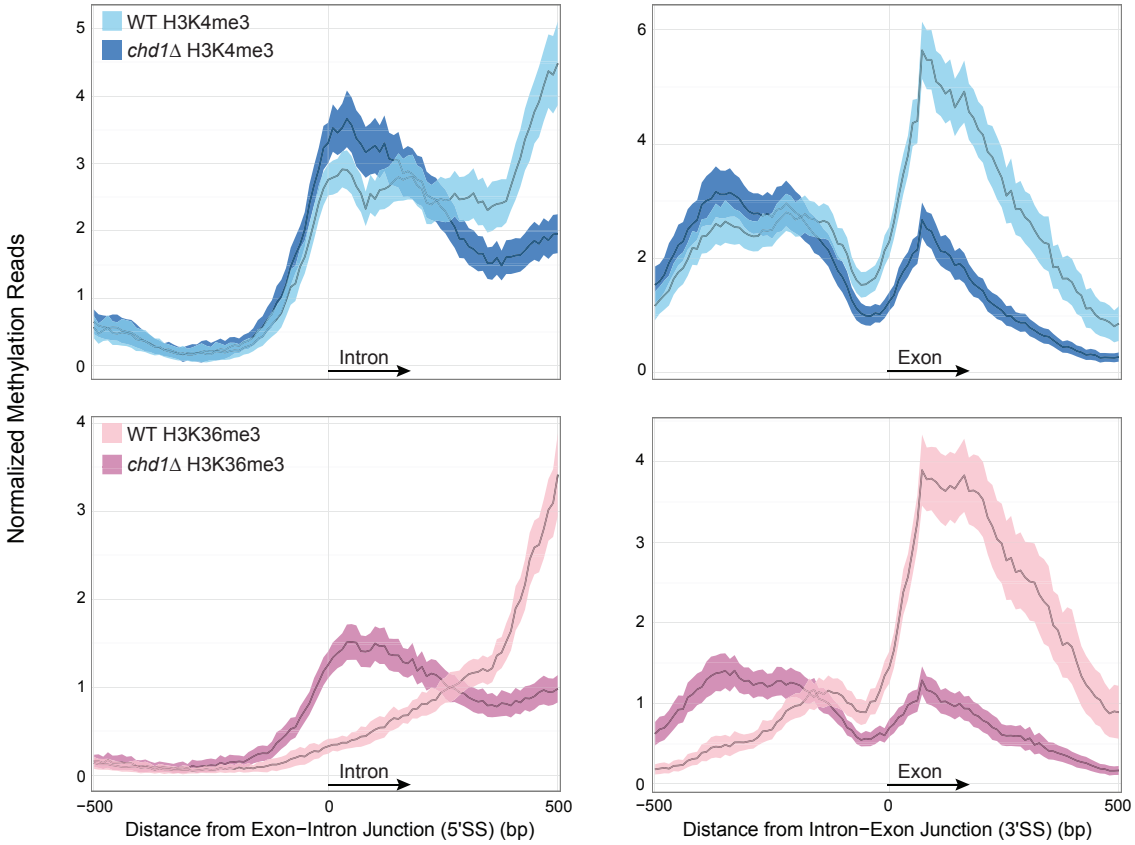


Figure 3.14: The H3K4me3 and H3K36me3 patterns at splice sites (SS) found within RP genes in the absence of *CHD1*. Average profiles of H3K4me3 (top) and H3K36me3 (bottom) at 5' SS (left) and 3' SS (right) for 101 RP genes having intron(s), corresponding to 104 introns, in *chd1Δ* and WT. H3K4me3 and H3K36me3 signals in WT are colored in sky blue and pink, respectively. H3K4me3 and H3K36me3 signals in *chd1Δ* are colored in dark blue and purple, respectively. The middle line indicates the mean of the normalized methylation reads for RP genes, and the shaded area around the line is a 95% confidence interval for the mean. The unit of the y-axis is RPM.

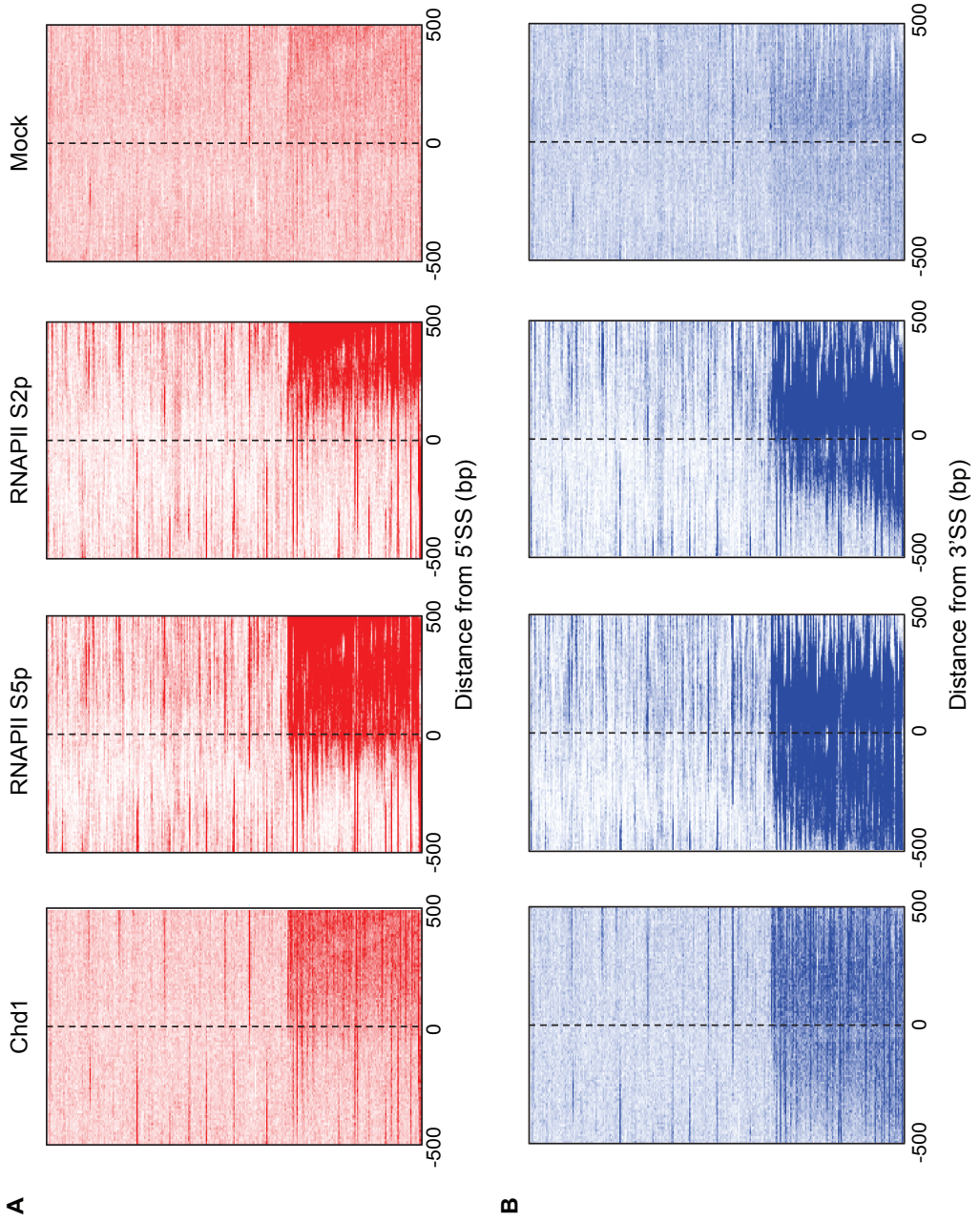


Figure 3.15: Chd1 occupancy at 5' splice sites (SS) (A) and at 3' SS (B) found within all introns[i] in yeast. For the reference of transcription level across the introns, RNAPII S5p and RNAPII S2p datasets are incorporated. For the negative control, Mock data is

[i] Introns within tRNA genes were excluded.

(Figure 3.15 continued.) added. Occupancies of Chd1, RNAPII S5p, RNAPII S2p, and Mock are measured by counting ChIP sequencing reads from 500 bp upstream to 500 bp downstream of 5' SS and of 3' SS. The rows were sorted by transcription level measured by RNAPII S5p.

To test this hypothesis, we performed RNA-sequencing of *chd1Δ* and WT and counted the intronic and coding exonic sequencing reads separately for all intron-containing genes. To compare intron retention between *chd1Δ* and WT, we calculated the odds ratio (OR) of intron retention for individual introns as follows:

$$\frac{(Number\ of\ reads\ mapped\ to\ intron/\ Number\ of\ reads\ mapped\ to\ CDS)\ in\ chd1\Delta}{(Number\ of\ reads\ mapped\ to\ intron/Number\ of\ reads\ mapped\ to\ CDS\ )\ in\ WT}$$

Essentially, this measurement informs us whether intron retention is higher or lower in RNA transcripts in the *CHD1* mutant compared to WT depending on whether the OR is greater or less than 1, respectively, in conjunction with a lenient p-*value* cutoff of < 0.1 from Fisher's exact test. Table 3.2 shows the outcome of the intron retention analysis for all intron-containing genes considered.

Table 3.2: The number of introns that are significantly different in intron retention between *chd1Δ* and WT, when all introns in the yeast genome were considered.

	In this study, Number of introns	Lee *et al*., (Lee et al., 2012) Number of introns
OR < 1, less intron retention in *chd1Δ*	28	98
OR > 1, more intron retention in *chd1Δ*	7	6

We found 35 introns were significantly affected in splicing in the Chd1 mutant; 28 out of 35 introns showed an improvement in splicing. We implemented this analysis

with more deeply sequenced RNA-sequencing data, which was published in a previous study (Lee et al., 2012), since we lost many introns due to a lack of statistical power. This revealed that the directionality was more pronounced (*i.e.*, 6 out of 104 intron).

To validate our findings, we applied this analysis to publicly available RNA-sequencing data generated with a splicing factor mutant (temperature-sensitive *prp40; prp40-1*) (Volanakis et al., 2013), and found the opposite trend – over 95% of introns were impaired in splicing (data not shown), suggesting our analysis was valid.

When we considered the introns within RP genes, where Chd1 is enriched, the increased efficiency in splicing (*i.e.*, less introns retained in transcribed mRNAs) was even more apparent (Table 3.3).

Table 3.3: The number of introns that are significantly different in intron retention between *chd1Δ* and WT, when the introns at RP genes were considered.

	In this study, Number of introns	Lee *et al.*, (Lee et al., 2012) Number of introns
OR < 1, less intron retention in *chd1Δ*	21	88
OR > 1, more intron retention in *chd1Δ*	2	1

This result suggests Chd1 is functionally linked with RNA splicing. Therefore, we concluded that Chd1 affects RNA splicing by most likely modulating rate of RNAPII elongation. Another compelling mechanism might be through interactions between

histone methylation and splicing machinery, as demonstrated in a recent study where loss of *SET2* recruits inadequate snRNPs to yeast chromatin (Sorenson et al., 2016).

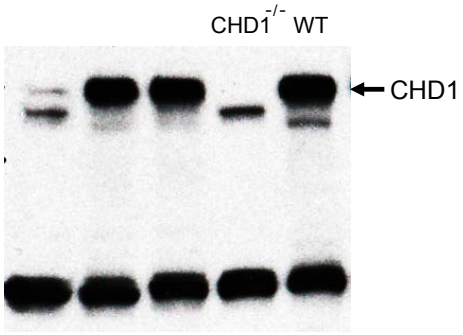
3.3.5 GENERATING *CHD1* KNOCK OUT IN HUMAN GLIOBLASTOMA CELL LINE VIA THE CRISPR/CAS9 GENOME EDITING SYSTEM

Although the role of *CHD1* has been recognized in cancer cells (Burkhardt et al., 2013; Huang et al., 2012; Liu et al., 2012; Rodrigues et al., 2015; Yao et al., 2015), its role in development and progress of glioblastoma multiforme (GBM) has not been studied yet. In order to aid in understanding the role of *CHD1* in GBM, we generated a knockout (KO) cell line of *CHD1* using the CRISPR (Clustered, Regularly Interspaced, Short Palindromic Repeat) system. The CRISPR system is a prokaryotic adaptive immune system that uses a RNA-guided DNA nuclease (*i.e.*, Cas9) to silence viral nucleic acids. The CRISPR/Cas9 system is widely used as an efficient and reliable way to make precise and targeted changes to the genome *in vivo* (Hsu et al., 2014). We took advantage of this tool to create *CHD1* KO in T98G GBM cell line.

Human *CHD1* is located on chromosome 5: 98,853,985 – 98,928,957. We used the longest transcript (CHD1-001, ENST00000614616) as our reference to determine the target locus for the CRISPR/Cas9 system on the genome. Because the start codon (*i.e.*, ATG) is located at the end of the second exon in this transcript, we designed the guide RNA for CRISPR/Cas9 to recognize the start of the third exon and introduce a nucleotide cut, leading to a premature stop codon. This exon is the common exon for two *CHD1*

transcripts (*i.e.*, ENST00000614616, ENST00000284049) that are assigned as a principal isoform by the APPRIS system (Rodriguez et al., 2013). We cloned the guide RNA into a vector with all functional elements needed for the CRSPR/Cas9 system (GeneArt CRISPR Nuclease Vector with CD4, invitrogen), and transfected the plasmid into T98G cells. We then isolated the cells expressing the reporter (*i.e.*, CD4) using magnetic beads coated with anti-human CD4 antibody (Dynabeads CD4 Positive Isolation Kit, Invitrogen). Subsequently, we isolated monoclones by seeding a limited number of cells onto a large culture dish, allowing the cells to grow sparsely to form discernible colonies. Then, we expanded individual colonies, and confirmed them by western blot and genomic sequencing once the cells were confluent, until we found a correct clone (Figure 3.16).

A



B

Query: *CHD1* Knockout (CHD1$^{-/-}$)
Subject: Human Reference Genome

```
Query  161  TGCTTCCATCACTACTGCTTCCAGAACTCGAACCAGATCCAGAGCCTGAAGCTGACCCAG  220
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  TGCTTCCATCACTACTGCTTCCAGAACTCGAACCAGATCCAGAGCCTGAAGCTGACCCAG  180

Query  221  AATCATCAT-CGACTGGCTATAATTTGAAAATAAACAATTTCAAACAAAATTCATTAGGA  279
            ||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  AATCATCATCCGACTGGCTATAATTTGAAAATAAACAATTTCAAACAAAATTCATTAGGA  240
```

Figure 3.16: Confirmation of *CHD1* knockout clone(s) created by the CRISPR/Cas9 system. (A) Confirmation western blot. The left three columns are incorrect *CHD1* knockout clones, the fourth column is a correct *CHD1* knockout (CHD1$^{-/-}$), and the fifth column is WT (*i.e.*, positive control). The bottom band on the western blot image was used for the loading control to estimate the amount of protein loaded on sample wells in a polyacrylamind gel. (B) Confirmation genomic DNA sequencing. The Query is the (Figure 3.16 continued.) genomic sequences obtained from the correct *CHD1* knockout verified by western blot. The Subject is the nucleotide sequences in human reference genome. The nucleotide sequences colored in yellow are what we initially targeted for the CRISPR/Cas9 system to recognize, and the red asterisk indicates a single nucleotide deletion introduced by CRISPR/Cas9, leading to a premature stop codon.

3.3.6 *CHD1* FUNCTION IN HUMAN GLIOBLASTOMA CELL LINE

With the *CHD1* knockout cell, we generated gene expression profiles to investigate transcriptional changes due to loss of *CHD1* through RNA sequencing. We

used Kallisto (Bray et al., 2016) to quantify abundance of transcripts in RNA sequencing data and validated expression level of *CHD1* in our *CHD1* knockout cell. We found that expression of *CHD1* transcripts was significantly lower in $CHD1^{-/-}$ cells than WT cells (Figure 3.17). We performed differential expression analysis by using Sleuth (Harold J Pimentel, 2016), and found that Chd1 KO affected expression of a considerable number of transcripts (Table 3.4).

To understand biological function of these differentially expressed genes, we conducted GO analysis with 404 activated genes (Table 3.5) and 470 repressed genes (Table 3.6) in Chd1 KO (adjusted p-value < 0.01 and $|b| > 1$) by using Enrichr (Chen et al., 2013). We found that the sterol metabolic process was highly enriched in the up-regulated genes while the extracellular matrix-related process including regulation of cell adhesion was top-ranked in the down-regulated genes in Chd1 KO. It would be interesting to find how a chromatin remodeler (*i.e.*, Chd1) specifically affects these biological processes in this cancer cell line.

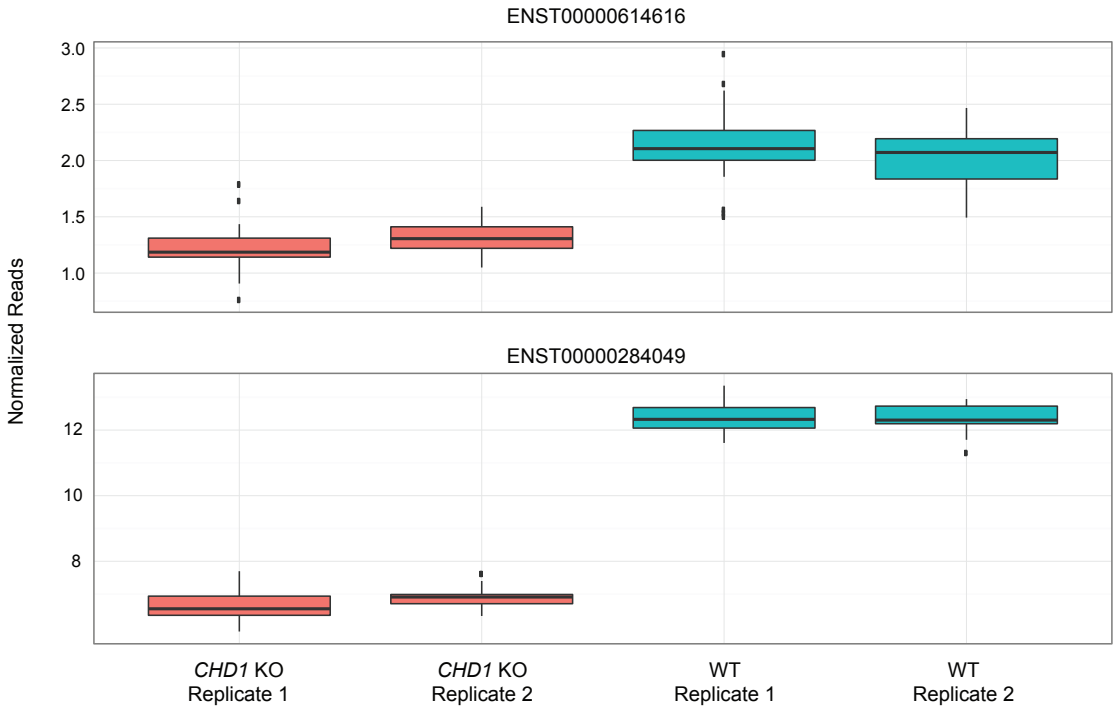


Figure 3.17: Expression of *CHD1* transcripts in *CHD1* KO cells. ENST00000614616 (top panel) and ENST00000284049 (bottom panel) are two major transcript isoforms generated from *CHD1* gene. *Notes* – The units for normalized reads on the y-axis are TPM (Number of transcripts per million).

Table 3.4: The number of transcripts that are differentially expressed in *CHD1* KO depending on varying thresholds. *Notes* – 'b' is the 'beta' value reported in the Sleuth output, which is the bias of an estimator. 'b' refers to the difference between estimator's expected value and the true value, which is analogous to fold change of Chd1 KO relative to WT in expression.

	Adjusted p-value < 0.01				
	$\|b\| > 0$	$\|b\| > 0.5$	$\|b\| > 1$	$\|b\| > 1.5$	$\|b\| > 2$
Down-regulation in *CHD1* KO	2,734	1,455	470	231	162
Up-regulation in *CHD1* KO	2,695	1,368	404	184	114

Table 3.5: GO biological process terms significantly enriched in activated genes in *CHD1* KO. The p-value and score are computed by Enrichr to assess the significance overlap between the input list of genes and the background list of genes in each GO term. The p-value is a result of Fisher's exact test and the score is obtained by multiplying the log of the p-value by z-score which represents deviation from the expected rank by Fisher's exact test (Chen et al., 2013).

GO Term	**P-value**	**Score**
Sterol biosynthetic process	5.09E-10	50.15
Cholesterol biosynthetic process	9.85E-10	50.01
Steroid biosynthetic process	4.18E-08	37.99
Alcohol biosynthetic process	1.28E-07	35.36
Sterol metabolic process	1.22E-07	35.14
Cholesterol metabolic process	1.56E-07	34.44
Organic hydroxy compound biosynthetic process	6.97E-06	27.05
Steroid metabolic process	8.31E-06	26.74
Small molecule biosynthetic process	2.04E-05	25.50
Lipid biosynthetic process	8.30E-05	22.19
Alcohol metabolic process	2.97E-04	18.86
Wound healing, spreading of cells	3.85E-03	15.18
Organic hydroxy compound metabolic process	3.85E-03	12.95
Isoprenoid biosynthetic process	1.58E-02	11.43
Regulation of response to wounding	1.03E-02	11.22
Regulation of homeostatic process	1.03E-02	11.10
Fatty acid biosynthetic process	8.91E-03	10.52
Positive regulation of cation channel activity	2.31E-02	10.10

Table 3.6: GO biological process terms significantly enriched in repressed genes in *CHD1* KO. The p-value and score are computed by Enrichr to assess the significance overlap between the input list of genes and the background list of genes in each GO term. The p-value is a result of Fisher's exact test and the score is obtained by multiplying the log of the p-value by z-score which represents deviation from the expected rank by Fisher's exact test (Chen et al., 2013)

GO Term	**P-value**	**Score**
Extracellular matrix organization	2.63E-05	25.11
Extracellular structure organization	2.63E-05	25.11
Calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules	4.37E-04	18.54
Tube development	5.04E-04	17.66
Morphogenesis of an epithelium	8.32E-04	16.61
Regulation of cell adhesion	1.21E-03	16.42
Regulation of chondrocyte differentiation	1.42E-03	16.40
Positive regulation of locomotion	1.83E-03	15.43
Positive regulation of cartilage development	3.24E-03	15.42
Positive regulation of cell migration	1.91E-03	15.34
Odontogenesis	8.32E-04	15.34
Regulation of cartilage development	1.36E-03	15.31
Organ morphogenesis	1.45E-03	15.28
Positive regulation of cell motility	2.14E-03	15.04
Regulation of endothelial cell proliferation	1.36E-03	15.00
Cell-substrate adhesion	1.11E-03	14.99
Tissue morphogenesis	1.83E-03	14.80
Positive regulation of cellular component movement	2.49E-03	14.67

(Table 3.6 continued.)

GO Term	P-value	Score
Positive regulation of epithelial cell proliferation	2.24E-03	14.21
Regulation of chemotaxis	2.14E-03	14.11
Organ induction	4.89E-03	13.83
Cell-marix adhesion	2.14E-03	13.70
Positive regulation of endothelial cell proliferation	2.14E-03	13.63
Positive regulation of behavior	2.83E-03	13.32
Kidney development	2.24E-03	13.26
Homophilic cell adhesion via plasma membrane adhesion molecules	2.14E-03	13.26
Regulation of vasculature development	2.83E-03	13.26
Integrin-mediated signaling pathway	2.37E-03	13.21
Regulation of organ formation	4.89E-03	13.01

Appendix A – List of Immunoprecipitants

This is a list of immunoprecipitants that we used in Chapter 2 and Chapter 3.

Immunoprecipitant	Product name	Product company
Anti-Tap Beads	IgG Sepharose 6 Fast Flow	GE Healthcare (#17-0969)
Anti-Myc Beads	EZview Red Anti-c-Myc Affinity Gel	Sigma (#E6654)
Anti-RNAPII CTD S5p	Anti-RNA polymerase II CTD repeat YSPTSPS (phospho S5) antibody	Abcam (#ab5131)
Anti-RNAPII CTD S2p	Anti-RNA polymerase II CTD repeat YSPTSPS (phospho S2) antibody	Abcam (#ab5095)
Anti-H3K4me3	Anti-trimethyl-Histone H3 (Lys4) antibody	EMD Millipore (#07-473)
Anti-H3K36me3	Anti-Histone H3 (tri methyl K36) antibody	Abcam (#ab9050)

Bibliography

Akerfelt, M., Morimoto, R.I., and Sistonen, L. (2010). Heat shock factors: integrators of cell stress, development and lifespan. Nature reviews Molecular cell biology *11*, 545-555.

Alexander, R.D., Innocente, S.A., Barrass, J.D., and Beggs, J.D. (2010). Splicing-dependent RNA polymerase pausing in yeast. Molecular cell *40*, 582-593.

Annunziato, A. (2008). DNA Packaging: Nucleosomes and Chromatin. Nature Education *1*, 26.

Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M.K., Chuang, R., Jaehnig, E.J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M.*, et al*. (2010). Rewiring of genetic networks in response to DNA damage. Science *330*, 1385-1389.

Biswas, D., Dutta-Biswas, R., and Stillman, D.J. (2007). Chd1 and yFACT act in opposition in regulating transcription. Molecular and cellular biology *27*, 6279-6287.

Black, J.C., Van Rechem, C., and Whetstine, J.R. (2012). Histone lysine methylation dynamics: establishment, regulation, and biological impact. Molecular cell *48*, 491-507.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nature biotechnology *34*, 525-527.

Burkhardt, L., Fuchs, S., Krohn, A., Masser, S., Mader, M., Kluth, M., Bachmann, F., Huland, H., Steuber, T., Graefen, M.*, et al*. (2013). CHD1 is a 5q21 tumor suppressor required for ERG rearrangement in prostate cancer. Cancer research *73*, 2795-2805.

Causton, H.C., Ren, B., Koh, S.S., Harbison, C.T., Kanin, E., Jennings, E.G., Lee, T.I., True, H.L., Lander, E.S., and Young, R.A. (2001). Remodeling of yeast genome

expression in response to environmental changes. Molecular biology of the cell *12*, 323-337.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC bioinformatics *14*, 128.

Clapier, C.R., and Cairns, B.R. (2009). The biology of chromatin remodeling complexes. Annual review of biochemistry *78*, 273-304.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America *95*, 14863-14868.

Fromm, G., Gilchrist, D.A., and Adelman, K. (2013). SnapShot: Transcription regulation: pausing. Cell *153*, 930-930 e931.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Molecular biology of the cell *11*, 4241-4257.

Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737-741.

Giardina, C., and Lis, J.T. (1995). Dynamic protein-DNA architecture of a yeast heat shock promoter. Molecular and cellular biology *15*, 2737-2744.

Gietz, R.D., and Schiestl, R.H. (2007). High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. Nature protocols *2*, 31-34.

Gkikopoulos, T., Schofield, P., Singh, V., Pinskaya, M., Mellor, J., Smolle, M., Workman, J.L., Barton, G.J., and Owen-Hughes, T. (2011). A role for Snf2-related nucleosome-spacing enzymes in genome-wide nucleosome organization. Science *333*, 1758-1760.

Hahn, J.S., Hu, Z., Thiele, D.J., and Iyer, V.R. (2004). Genome-wide analysis of the biology of stress responses through heat shock transcription factor. Molecular and cellular biology *24*, 5249-5256.

Hajheidari, M., Koncz, C., and Eick, D. (2013). Emerging roles for RNA polymerase II CTD in Arabidopsis. Trends in plant science *18*, 633-643.

Hall, D.A., Zhu, H., Zhu, X., Royce, T., Gerstein, M., and Snyder, M. (2004). Regulation of gene expression by a metabolic enzyme. Science *306*, 482-484.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J.*, et al*. (2004). Transcriptional regulatory code of a eukaryotic genome. Nature *431*, 99-104.

Harold J Pimentel, N.B., Suzette Puente, Páll Melsted, Lior Pachter (2016). Differential analysis of RNA-Seq incorporating quantification uncertainty. bioRxiv.

Hsu, P.D., Lander, E.S., and Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. Cell *157*, 1262-1278.

Hu, Z., Killion, P.J., and Iyer, V.R. (2007). Genetic reconstruction of a functional transcriptional regulatory network. Nature genetics *39*, 683-687.

Huang, S., Gulzar, Z.G., Salari, K., Lapointe, J., Brooks, J.D., and Pollack, J.R. (2012). Recurrent deletion of CHD1 in prostate cancer with relevance to cell invasiveness. Oncogene *31*, 4164-4170.

Jakobsen, B.K., and Pelham, H.R. (1988). Constitutive binding of yeast heat shock factor to DNA in vivo. Molecular and cellular biology *8*, 5040-5042.

Jonkers, I., and Lis, J.T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. Nature reviews Molecular cell biology *16*, 167-177.

Jothi, R., Balaji, S., Wuster, A., Grochow, J.A., Gsponer, J., Przytycka, T.M., Aravind, L., and Babu, M.M. (2009). Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. Molecular systems biology *5*, 294.

Kim, H., Erickson, B., Luo, W., Seward, D., Graber, J.H., Pollock, D.D., Megee, P.C., and Bentley, D.L. (2010). Gene-specific RNA polymerase II phosphorylation and the CTD code. Nature structural & molecular biology *17*, 1279-1286.

Kim, S., Bugga, L., Hong, E.S., Zabinsky, R., Edwards, R.G., Deodhar, P.A., and Armstrong, J.A. (2016). An RNAi-Based Candidate Screen for Modifiers of the CHD1 Chromatin Remodeler and Assembly Factor in Drosophila melanogaster. G3 (Bethesda) *6*, 245-254.

Kim, T., and Buratowski, S. (2009). Dimethylation of H3K4 by Set1 recruits the Set3 histone deacetylase complex to 5' transcribed regions. Cell *137*, 259-272.

Kim, T., Xu, Z., Clauder-Munster, S., Steinmetz, L.M., and Buratowski, S. (2012). Set3 HDAC mediates effects of overlapping noncoding transcription on gene induction kinetics. Cell *150*, 1158-1169.

Krogan, N.J., Kim, M., Tong, A., Golshani, A., Cagney, G., Canadien, V., Richards, D.P., Beattie, B.K., Emili, A., Boone, C., *et al*. (2003). Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. Molecular and cellular biology *23*, 4207-4218.

Kurdistani, S.K., and Grunstein, M. (2003). Histone acetylation and deacetylation in yeast. Nature reviews Molecular cell biology *4*, 276-284.

Lee, J.S., Garrett, A.S., Yen, K., Takahashi, Y.H., Hu, D., Jackson, J., Seidel, C., Pugh, B.F., and Shilatifard, A. (2012). Codependency of H2B monoubiquitination and nucleosome reassembly on Chd1. Genes & development *26*, 914-919.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., *et al*. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science *298*, 799-804.

Lenstra, T.L., Benschop, J.J., Kim, T., Schulze, J.M., Brabers, N.A., Margaritis, T., van de Pasch, L.A., van Heesch, S.A., Brok, M.O., Groot Koerkamp, M.J., *et al*. (2011). The specificity and topology of chromatin interaction pathways in yeast. Molecular cell *42*, 536-549.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Lin, J.J., Lehmann, L.W., Bonora, G., Sridharan, R., Vashisht, A.A., Tran, N., Plath, K., Wohlschlegel, J.A., and Carey, M. (2011). Mediator coordinates PIC assembly with recruitment of CHD1. Genes & development *25*, 2198-2209.

Liu, W., Lindberg, J., Sui, G., Luo, J., Egevad, L., Li, T., Xie, C., Wan, M., Kim, S.T., Wang, Z., *et al*. (2012). Identification of novel CHD1-associated collaborative alterations

of genomic structure and functional assessment of CHD1 in prostate cancer. Oncogene *31*, 3939-3948.

Liu, X., Kraus, W.L., and Bai, X. (2015). Ready, pause, go: regulation of RNA polymerase II pausing and release by cellular signaling pathways. Trends in biochemical sciences *40*, 516-525.

Lun, A.T., and Smyth, G.K. (2016). csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. Nucleic acids research *44*, e45.

Martin, C., and Zhang, Y. (2005). The diverse functions of histone lysine methylation. Nature reviews Molecular cell biology *6*, 838-849.

Nadal-Ribelles, M., Mas, G., Millan-Zambrano, G., Sole, C., Ammerer, G., Chavez, S., Posas, F., and de Nadal, E. (2015). H3K4 monomethylation dictates nucleosome dynamics and chromatin remodeling at stress-responsive genes. Nucleic acids research *43*, 4937-4949.

Narlikar, G.J., Sundaramoorthy, R., and Owen-Hughes, T. (2013). Mechanisms and functions of ATP-dependent chromatin-remodeling enzymes. Cell *154*, 490-503.

Nourani, A., Robert, F., and Winston, F. (2006). Evidence that Spt2/Sin1, an HMG-like factor, plays roles in transcription elongation, chromatin structure, and genome stability in Saccharomyces cerevisiae. Molecular and cellular biology *26*, 1496-1509.

Park, D., Shivram, H., and Iyer, V.R. (2014). Chd1 co-localizes with early transcription elongation factors independently of H3K36 methylation and releases stalled RNA polymerase II at introns. Epigenetics & chromatin *7*, 32.

Park, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nature reviews Genetics *10*, 669-680.

Petty, E., and Pillus, L. (2013). Balancing chromatin remodeling and histone modifications in transcription. Trends in genetics : TIG *29*, 621-629.

Phatnani, H.P., and Greenleaf, A.L. (2006). Phosphorylation and functions of the RNA polymerase II CTD. Genes & development *20*, 2922-2936.

Quan, T.K., and Hartzog, G.A. (2010). Histone H3K4 and K36 methylation, Chd1 and Rpd3S oppose the functions of Saccharomyces cerevisiae Spt4-Spt5 in transcription. Genetics *184*, 321-334.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Radman-Livaja, M., Quan, T.K., Valenzuela, L., Armstrong, J.A., van Welsem, T., Kim, T., Lee, L.J., Buratowski, S., van Leeuwen, F., Rando, O.J., *et al*. (2012). A key role for Chd1 in histone H3 dynamics at the 3' ends of long genes in yeast. PLoS genetics *8*, e1002811.

Richter, K., Haslbeck, M., and Buchner, J. (2010). The heat shock response: life on the verge of death. Molecular cell *40*, 253-266.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research *43*, e47.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

Rodrigues, L.U., Rider, L., Nieto, C., Romero, L., Karimpour-Fard, A., Loda, M., Lucia, M.S., Wu, M., Shi, L., Cimic, A., *et al*. (2015). Coordinate loss of MAP3K7 and CHD1 promotes aggressive prostate cancer. Cancer research *75*, 1021-1034.

Rodriguez, J.M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J.J., Lopez, G., Valencia, A., and Tress, M.L. (2013). APPRIS: annotation of principal and alternative splice isoforms. Nucleic acids research *41*, D110-117.

Schweikert, G., Cseke, B., Clouaire, T., Bird, A., and Sanguinetti, G. (2013). MMDiff: quantitative testing for shape changes in ChIP-Seq data sets. BMC genomics *14*, 826.

Shilatifard, A. (2012). The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. Annual review of biochemistry *81*, 65-95.

Simic, R., Lindstrom, D.L., Tran, H.G., Roinick, K.L., Costa, P.J., Johnson, A.D., Hartzog, G.A., and Arndt, K.M. (2003). Chromatin remodeling protein Chd1 interacts with transcription elongation factors and localizes to transcribed genes. The EMBO journal *22*, 1846-1856.

Sims, R.J., 3rd, Chen, C.F., Santos-Rosa, H., Kouzarides, T., Patel, S.S., and Reinberg, D. (2005). Human but not yeast CHD1 binds directly and selectively to histone H3 methylated at lysine 4 via its tandem chromodomains. The Journal of biological chemistry *280*, 41789-41792.

Sims, R.J., 3rd, Millhouse, S., Chen, C.F., Lewis, B.A., Erdjument-Bromage, H., Tempst, P., Manley, J.L., and Reinberg, D. (2007). Recognition of trimethylated histone H3 lysine

4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. Molecular cell *28*, 665-676.

Skene, P.J., Hernandez, A.E., Groudine, M., and Henikoff, S. (2014). The nucleosomal barrier to promoter escape by RNA polymerase II is overcome by the chromatin remodeler Chd1. eLife *3*, e02042.

Smolle, M., Venkatesh, S., Gogol, M.M., Li, H., Zhang, Y., Florens, L., Washburn, M.P., and Workman, J.L. (2012). Chromatin remodelers Isw1 and Chd1 maintain chromatin structure during transcription by preventing histone exchange. Nature structural & molecular biology *19*, 884-892.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology *3*, Article3.

Sorenson, M.R., Jha, D.K., Ucles, S.A., Flood, D.M., Strahl, B.D., Stevens, S.W., and Kress, T.L. (2016). Histone H3K36 methylation regulates pre-mRNA splicing in Saccharomyces cerevisiae. RNA biology *13*, 412-426.

Venters, B.J., Wachi, S., Mavrich, T.N., Andersen, B.E., Jena, P., Sinnamon, A.J., Jain, P., Rolleri, N.S., Jiang, C., Hemeryck-Walsh, C., *et al*. (2011). A comprehensive genomic binding map of gene and chromatin regulatory proteins in Saccharomyces. Molecular cell *41*, 480-492.

Volanakis, A., Passoni, M., Hector, R.D., Shah, S., Kilchert, C., Granneman, S., and Vasiljeva, L. (2013). Spliceosome-mediated decay (SMD) regulates expression of nonintronic genes in budding yeast. Genes & development *27*, 2025-2038.

Wagner, E.J., and Carpenter, P.B. (2012). Understanding the language of Lys36 methylation at histone H3. Nature reviews Molecular cell biology *13*, 115-126.

Westerheide, S.D., and Morimoto, R.I. (2005). Heat shock response modulators as therapeutic tools for diseases of protein conformation. The Journal of biological chemistry *280*, 33097-33100.

Wiederrecht, G., Seto, D., and Parker, C.S. (1988). Isolation of the gene encoding the S. cerevisiae heat shock transcription factor. Cell *54*, 841-853.

Workman, C.T., Mak, H.C., McCuine, S., Tagne, J.B., Agarwal, M., Ozier, O., Begley, T.J., Samson, L.D., and Ideker, T. (2006). A systems approach to mapping DNA damage response pathways. Science *312*, 1054-1059.

Xiao, T., Shibata, Y., Rao, B., Laribee, R.N., O'Rourke, R., Buck, M.J., Greenblatt, J.F., Krogan, N.J., Lieb, J.D., and Strahl, B.D. (2007). The RNA polymerase II kinase Ctk1 regulates positioning of a 5' histone methylation boundary along genes. Molecular and cellular biology *27*, 721-731.

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Munster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L.M. (2009). Bidirectional promoters generate pervasive transcription in yeast. Nature *457*, 1033-1037.

Yao, H., Pan, J., Wu, C., Shen, H., Xie, J., Wang, Q., Wen, L., Wang, Q., Ma, L., Wu, L., *et al*. (2015). Transcriptome sequencing reveals CHD1 as a novel fusion partner of RUNX1 in acute myeloid leukemia with t(5;21)(q21;q22). Molecular cancer *14*, 81.

Zanton, S.J., and Pugh, B.F. (2004). Changes in genomewide occupancy of core transcriptional regulators during heat stress. Proceedings of the National Academy of Sciences of the United States of America *101*, 16843-16848.

Zhang, L., Schroeder, S., Fong, N., and Bentley, D.L. (2005). Altered nucleosome occupancy and histone H3K4 methylation in response to 'transcriptional stress'. The EMBO journal *24*, 2379-2390.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., *et al*. (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology *9*, R137.

Vita

Yaelim Lee received her B.S and M.S in Life Sciences and Biotechnology from Korea University in 2005 and 2008, respectively. In the fall of 2009, she started her Ph.D. degree in the Cellular and Molecular Biology graduate program in the University of Texas at Austin.

Permanent address (or email): yaelim242@gmail.com

This dissertation was typed by Yaelim Lee.